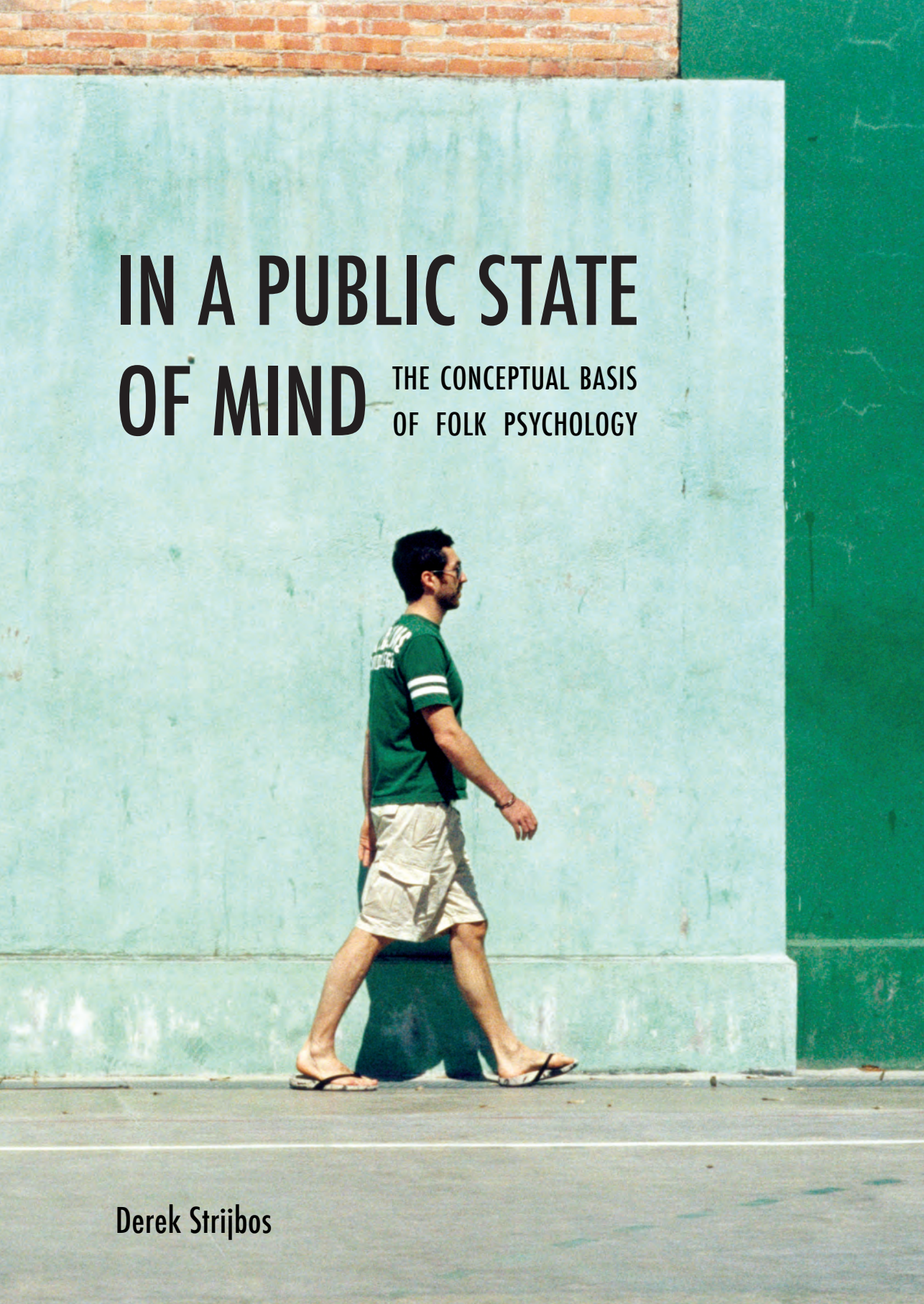
IN A PUBLIC STATE OF MIND

THE CONCEPTUAL BASIS OF FOLK PSYCHOLOGY



Derek Strijbos

In a Public State of Mind

The Conceptual Basis of Folk Psychology

Derek Strijbos

Cover photo by Luke Jarvis
Cover design by Bram Strijbos
ISBN/EAN: 9789090271194
Printed by Ipskamp Drukkers, Enschede

In a Public State of Mind

The Conceptual Basis of Folk Psychology

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus
prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op
donderdag 8 november 2012
om 15.30 uur precies

door

Derek Willem Strijbos
geboren op 24 augustus 1979
te Nijmegen

Promotor:
prof. dr. M.V.P. Slors

Manuscriptcommissie:
prof. dr. J.A.M. Bransen
prof. dr. G. Glas (Vrije Universiteit Amsterdam)
prof. dr. R. A. van der Sandt
prof. dr. L.B.W. Geurts
dr. L.C. de Bruin

Contents

Preface and Acknowledgements

When I started my PhD project, I was pretty sure I knew what my dissertation was going to be about. Looking back, five years later, I must conclude that I didn't have a clue at the time. I still don't quite understand how it happened, but the topic of my dissertation changed completely over the years. Inspired by my work in psychiatry, I started out with an idea about the nature of mental illness. I ended up writing a dissertation about the nature of commonsense psychology. From psychopathology to common sense: a rather surprising and admittedly radical shift of focus.

Now that I have started working with psychiatric patients again, I am slowly but surely starting to see the connection. The other day, I interviewed a patient who was plagued by compulsive doubt. Everything he was aware of thinking or doing had to face a tribunal of critical thought. This process of unrelenting self-reflection absorbed his mind, to the point that he felt mentally paralyzed, afraid to do or think anything at all.

As in so many cases in psychiatry, the story of this patient reveals a core feature of mental health, a feature which tends to remain unnoticed due to the very fact that it enables healthy mental activity. In this case, it has to do with our capacity to think, plan and perform our actions spontaneously, without a moment of reflection disturbing the flow of mind. What this patient had lost, I concluded after the interview, was the certainty of everyday practical life, the

certainty of common sense that makes up the necessary background against which productive thought and action becomes possible. Due to some basic affective disturbance, I figured, this person had turned into an overreflective self-interpreter who could no longer rely on his common sense, which put every spontaneous impulse of mental activity to a halt and left him practically incapacitated.

At the same time, I realized that this patient's incapacity to rely on common sense had had a profound influence on *my* use of common sense during the interview. In order to make this person in front of me feel comfortable enough to allow me to explore the world from his perspective, and to get an idea of what the world was like for him, I had actively *constrained* the ways I normally rely on common sense for the period of the interview. I had inhibited some of my spontaneous ways of thinking and interacting in order to make room for his discrepant personal perspective. At the level of social interaction too, the interview with this patient brought out the absence of something we normally take for granted: that in socially engaging with one another in everyday life, we unreflectively rely on our common sense, our common ways of experiencing, thinking and acting.

Reconstructing the development of my ideas during the last five years, it seems I shifted focus from the nature of psychiatric illness to the nature of the health care professional's *interaction with* people suffering from psychiatric illness, only to become fascinated by something which is significantly *impaired* in such interactions. This dissertation is about our reliance on common sense in everyday social engagements with one another. It argues against the consensus view on social cognition in philosophy and psychology, which, as it turns out, models our ordinary interactions with one another on the kind of situation faced by psychiatrists when interviewing their patients. On this view, the difference between the two is of a mere quantitative nature, the clinician simply having to account for *more* discrepant mental states than the ordinary interpreter in order to make sense of the personal perspective of the individual she is facing. The more I interact with patients, the more I am convinced that this view is deeply mistaken. Looking back, this dissertation can be seen as an attempt to account for this difference in a qualitative way. It tries to explain our reliance on common sense in social interaction in terms of the ascription of a kind of mental state that is *different in nature* from the kind we attribute when attempting to understand the discrepant views of other people, as in the case of interpreting psychiatric patients. Commonsense psychology, I argue, is based on the ascription of public mental states, states that relate the individuals we interact with to the world of common understanding.

Writing this dissertation often felt like a ride on an academic rollercoaster, with alternating feelings of philosophical excitement and intellectual nausea. It was a true adventure I wouldn't have missed for the world. There are many people who inspired, encouraged and supported me along the way. I wish to mention a few of them here.

First of all, I want to express my enormous gratitude to Marc Slors, who gave me the opportunity to start and complete this adventure. He has been an inspiring tutor, who gave me all the freedom I needed to find my own way into philosophy. Marc has the exceptional quality to tell you what your question is when you have already found the answer. This helped me tremendously to gain and regain focus in the process of writing this dissertation. I also want to thank Chris Buskes, my 'roommate' at the university, for his steady stream of encouragement, and for ignoring me when I was mumbling to myself or pacing the room, trying to figure out what I was thinking. Over the years I presented many drafts of my work at the Nijmegen Research Seminar Mind and Action. Thanks to all the participants who took the trouble to 'co-cognize' during my presentations and to comment on my writings, Fleur Jongepier, Bas Donders and Caroline Hoving-Boot in particular. While writing my dissertation, I had the privilege to work together with Leon de Bruin on several papers. Our collaboration has been a truly inspiring and fruitful mixture of talent and character, which I hope we can extend into the future. My days as a PhD-student wouldn't have been the same without the endless philosophical discussions with Sander Voerman. He deserves to be mentioned here, if only for never agreeing with me on any philosophical topic.

I am grateful to the organizers of several workshops and conferences, who gave me the opportunity to share my thoughts with a wider philosophical audience. In particular I want to thank Dan Hutto, who was so kind to invite me on several occasions. His generous hospitality and relentless philosophical energy have been a great source of inspiration. I would like to acknowledge the International Office of the Radboud University here, for providing me with the financial support to make these conference visits. Thanks also to the former Katholieke Radboud Stichting, now Stichting Thomas More, for giving me a grant to start exploring philosophy back in 2001. Without this grant, philosophy would probably have remained a distant and mysterious dream for me.

When I started my residency training in psychiatry last year, I hadn't finished my manuscript yet. I am very grateful to my supervisor Gerrit Glas for giving me the time and opportunity to write the last chapters during my first rotation last winter, and to Alexandâr Alexandrov, Aline Plinsinga, Cees Baas and Corina Capalneanu, for putting up with my absence during this period.

I wouldn't be where I am now without the continuous love and support from my closest friends and family. I want to thank them for making me realize, time and time again, how priviliged I am to be part of their lives. A special word of thanks to my oldest buddies Cyril Moers, Roel Schutgens and Jules van Binsbergen for accepting the honorable job of paranymph at the public defense of my dissertation, to my brother Bram for designing the cover of this book and for helping me with the layout, and to my sister Liesbeth for always lending a sympathetic ear during our many lunches together on campus the last few years. My parents, Terri and Willem, I owe more than I could possibly remember. But the weeks I spent with them last winter while finishing this book remain vivid in my memory. Thank you for your good care and company during this intensive period.

Süleyman and Isaak, my little boys, you two have enriched my life in ways I couldn't have imagined possible. Thank you for letting me share in your discoveries during the day and for sleeping at night!

My dear Ayse, a life with you turns out to be the best of all worlds for me. You make me happy. How you manage to do this, I still don't know, but I hope you never stop.

1

Introduction

1.1 The Reflective Fallacy

We are all familiar with projectionist fallacies. Consider anthropomorphism, the projection of human characteristics onto members of other species. Or take 'the curse of knowledge', the widely studied phenomenon of projecting one's own knowledge onto the more naïve person. Typically, such projectionist inclinations are countered by reflection. Adopting a critical stance toward the act of attribution often reveals the differences between oneself and the animal or other person, resulting in attenuation or withdrawal of the attributions made.

In this book I want to introduce a kind of projectionist fallacy that is different in this respect. Rather than being countered by taking a step back and reflecting on the matter, it precisely *arises out of* such reflection. In particular, it springs from *philosophical* reflection. This is what I call the 'reflective fallacy'. The reflective fallacy is the fallacy of projecting certain philosophical analyses of thought and action onto our commonsense conception of ourselves as thinking agents. It may be a quite general phenomenon in philosophy. Here, however, the focus lies on a specific instance: that of regarding a particularly reflective form of action understanding in terms beliefs and desires as an ad-

equate account of our spontaneous understanding of each other's actions in terms of goals and reasons.

It has become customary among philosophers and psychologists to use the term 'folk psychology' (or 'commonsense psychology') to refer to the rich set of concepts we employ in our everyday lives to make sense of each other's thoughts, feelings, utterances and actions. Within this conceptual framework, a central place is reserved for the concept of intentional action. It is the concept we deploy when making sense of each other's behavior in terms of goals and reasons for action. It applies to Bill, who explains that he couldn't come to the party yesterday *because* he had to finish his report *in order to* ensure his candidacy for the position that has just become vacant at work. Or to Betty, who, *considering* her father's sudden illness, decided to cancel her trip *so that* she could help him around the house. Making sense of others in terms of the goals and reasons that motivate them to make certain decisions and perform certain actions is a core activity of human social cognition. It reveals the human mind as a rational mind, a mind capable of reasoning about what to do and why, i.e. as a *discursive* mind.

Established wisdom in many corners of philosophy has it that our commonsense understanding of rational, discursive minds evolves around the concepts of belief and desire. On this picture, folk psychology is essentially a form of belief-desire psychology. The idea seems simple enough. To adopt a goal, one must have a desire to achieve something. And to perform a goal-directed action in response, one must have certain beliefs about the means to achieve it. Interpreting someone as acting intentionally must therefore involve deployment of the concepts of belief and desire. Thus, Bill must have had a *desire* to be eligible for the new job at work and must have *believed* that finishing his report in time would make him eligible. This, in turn, must have evoked his *desire* to finish his report yesterday evening, which, guided by his *belief* that he would not be able to do so if he went to the party, resulted in his decision to stay home working on his report.

But there is more. Belief-desire psychology works irrespectively of the feasibility, truth or appropriateness of the beliefs and desires ascribed. Little Peter wanted to fly to the moon so he made himself a couple of wings from old newspapers. Cathy decided to buy a lottery ticket because she was convinced she was going to win the jackpot this time. Of course little Peter wouldn't fly to the moon in a million years, and the chances of Cathy winning the lottery were just about as slim. Belief-desire psychology enables us to make sense of all forms of intentional action with the use of one single explanatory strategy, whether the actions under consideration are realistic, acceptable, appropriate,

or not. And it makes sense to everyone. Just take a moment and reflect on the last thing you did before picking up this book. No doubt you can find an explanation that fits the scheme of belief-desire psychology. It's really amazing. It works every time.

As coherent and commonsensical as this story may sound, I think it presents us with a deeply distorted picture of our ordinary understanding of one another as discursively engaged human beings who perform goal-directed actions for reasons. It exemplifies a relatively detached, typically philosophical way of thinking about mind and action, which is not representative for our commonsense conception of discursive agency. Belief-desire psychology, so I will argue, is not the conceptual core of folk psychology; it only appears to be from a particularly reflective stance. If true, this claim is not only of philosophical significance. Following the philosophical orthodoxy, many psychologists and other cognitive (neuro-)scientists these days take belief-desire psychology as their starting point for further empirical inquiry, as the central *explanandum* of folk psychology their theories are designed to explain. If I am correct, however, they have been targeting the wrong phenomenon, mistaking what typically serves as a philosophical reconstruction *of* social cognition for the actual psychology of human discursive engagement.

The problem with this so-called 'Belief-Desire Model' of folk psychology is that it portrays our sensitivity to the discursive minds of others in an exclusively representationalist and individualist, subjectivist way. Beliefs and desires are representational states. As Davidson (1983/2001c, p. 138) once observed: "Much of the point of the concept of belief is the potential gap it introduces between what is held to be true and what is true." Mastery of the concept of belief requires that we be able to ascribe *false* beliefs, informational states that misrepresent the world. Likewise, mastery of the concept of desire is evidenced by the capacity to ascribe unrealistic, unacceptable or conflicting desires, motivational states that misrepresent the world as we ourselves want or expect it to be. Ascription of belief and desire *as such* must go accompanied by an acknowledgement of the possibility that the ascribed beliefs and desires are or turn out to be false, unfulfilled or inappropriate. Modeling our commonsense understanding of intentional action exclusively on these concepts, the Belief-Desire Model thus gives a thoroughly individualist picture of the folk psychological conception of mind. It depicts our understanding of the relation between another agent and the world as being mediated by her (mis)representations *of*, her subjective and possibly inaccurate or inappropriate views *on* the world. On the Belief-Desire Model, we perceive the discursive minds of others as essentially *private* minds. Accordingly, interpreting someone in terms of her

beliefs and desires reveals the way *she* (mis)takes the world to be, and not necessarily anyone else.

This picture flies in the face of common sense, or so I will argue. The main aim of this book is to show why this is so and to put an alternative picture in its place. I will call it the 'Relational Model' of folk psychology. Against the representationalist aspect of current orthodoxy, I will argue that interpretation of others in terms of their goals and reasons is first and foremost a kind of *relational* sense-making. Our default mode of discursive understanding of another person consists in quite literally seeing a connection between that person and her goals and reasons, as if by drawing an arrow from the person to certain salient and significant (past, present or future) events or situations. In such cases, interpreting someone starts with looking out into the world in search of her goals and reasons and it ends with finding them there. At no time during this interpretative act do we conceive of the other person as a world-representer.

Against the individualist or subjectivist aspect, I hold that interpreting others in terms of their goals and reasons takes place against the background of the *common* world. We tend to treat each other's intentional attitudes as *intersubjective* phenomena. This should not be understood as claiming that we interpret another person's intentional attitudes through, or as, sharing those attitudes ourselves, nor as saying that his or her intentional attitudes constitutively depend on our own presence or social engagement. Both may be true in some cases. We sometimes do instantiate the same intentional attitudes when attending or acting jointly, and, more controversially, an interpretee's intentional attitude towards an interpreter may be partly constituted (rather than merely caused) by the interpreter's presence and engagement. The claim here is rather this: that in our daily social affairs, we generally assume a person's intentional attitudes to consist in a relation between the *individual* person and the *common* world. This turns the minds of others into essentially *public* entities, entities whose acts of thinking and intending imply the existence of a community to which they belong. On this alternative picture of folk psychology, our basic understanding of the thoughts and actions of others is confined to how they ought to think and act under the circumstances, in accordance with established socio-cultural norms of reasoning and proper conduct. The private, representationalist conception of mind has no role to play in our spontaneous folk psychological practices.

I do not aspire to be a radical eliminativist about representationalist belief-desire psychology, however. Ascribing private beliefs and desires does belong to our folk psychological repertoire. But it is *not* our primary way of discursively engaging with others, not even when it comes to attributing goals and reasons

for action. Belief-desire psychology, I claim, is an essentially *complementary* interpretation strategy. It is not the conceptual core of folk psychology, nor the driving psychological force behind our discursive engagements with one another. Rather, it is an interpretative tool designed for reflection *on* and management *of* such discursive engagements when our default, relational modus of understanding runs aground.

Herein also lies the source of the reflective fallacy. As I pointed out above, it is amazing that belief-desire psychology always works. Parsing goals and reasons in terms of beliefs and desires is a move that is always available for a competent participant of discursive practice. Any story about folk psychology should be able to account for this remarkable fact. The line of reasoning that leads to the reflective fallacy starts with an acknowledgement of this fact, but it then derails by trying to explain it in terms of the *conceptual* structure of folk psychology or the *psychological* requirements for wielding it. The alternative is to account for this fact in terms of the *social* function of belief-desire psychology in human social life.

1.2 Preview

The distinction between the private and the public dimension of the commonsense mind has generally been neglected in the debate on folk psychology. Getting it into full view requires that we dig deeper into the conceptual framework in which much of the debate has been taking place. The private-public distinction does not parallel the oppositions that have held the debate in its grip over the last 30 years. It runs orthogonal to the question whether deployment of folk psychology is primarily a matter of theorizing about one another by means of a 'theory of mind', as the 'Theory Theory' of folk psychology has it, or rather consists in a simulative procedure during which we try to understand the world from the other person's point of view, as on the 'Simulation Theory' (e.g. Davies and Stone 1995a, 1995b, Carruthers and Smith 1996). Nor does it match the distinction between an 'internal' conception of mind on the one hand and an 'enacted' conception on the other. On the first, social cognition is best explained as a process of getting access to the mental states behind the expressions we see and the utterances we hear, entities that lie underneath the surface of social interaction, to be inferred from the outside by means of a theory or to be experienced from the inside by means of simulation (e.g. Herschbach 2008b, Spaulding 2010, Jacob 2011). On the second, other minds are encountered as essentially embodied entities as they take shape in the in-

teraction itself, entities directly perceived in the behavior displayed, whose boundaries are drawn only by the activities they engage in (e.g. Gallagher 2011, Zahavi 2011, Hutto 2011c). As we shall see in the following chapters, all these different and sometimes mutually incompatible accounts *can* and *should* acknowledge the folk psychological distinction between the private and the public sphere of mind.

Chapter 2 starts with a discussion of the Belief-Desire Model of action explanation as it has been developed in action theory and shows how this account has been widely adopted in the debate on folk psychology. It also gives a first impression of the Relational Model and introduces *relational mindreading* as a technical notion for the relational understanding of other people's goals and reasons for action in everyday discursive practice. Relational mindreading is a form of social understanding through which we perceive others as being intentionally directed toward the world in propositionally articulated, truth-evaluable ways, *without*, however, conceiving of them as *representing* the world in those ways. Relational mindreading thus consists in the ascription of *non*-representational, relational propositional attitudes.

To many a philosopher's ear, talking about a mental state with propositional content *just is* talking about a representational state. Accordingly, conceiving of someone as *saying* or *thinking* that such-and-such entails appreciation of the fact that he or she *represents* the world as such-and-such in doing so, or at least presupposes this fact. From this perspective, the idea of a non-representationalist folk conception of the propositional attitudes is simply incoherent. I take this orthodoxy seriously and will therefore consider it a genuine challenge for the Relational Model to reveal the notion of a non-representational propositional attitude as conceptually coherent, i.e. to establish the validity of the conceptual distinction between (the attribution of) such relational propositional states and their representational counterparts.

This first challenge for the Relational Model sets the agenda for chapters 3 and 4. Chapter 3 provides an elaborate discussion of Sellars's well-known 'Myth of Jones' in his 'Empiricism and the Philosophy of Mind' (1956/1997). Sellars myth has often been considered the intellectual inspiration for representationalist Theory Theories of folk psychology. But it actually turns out to be the perfect philosophical tool for introducing a mere *relational* conception of mindreading. Chapter 4 goes beyond Sellars's particular conceptual framework and shows how we can make the distinction between the representational/private and the relational/public dimension of mind irrespective of specific commitments regarding the nature of mental states and mental state attribution. We can cash out the distinction between relational and re-

presentational mindreading on *all* dominant theories in the literature: Theory Theories, Simulation Theories, internalist and enactivist accounts alike

Having met the challenge of conceptual validity, the next question is whether we should expect the distinction between (the ascription of) relational and representational propositional attitudes to make an actual difference in human social affairs. Besides being conceptually valid, is the distinction also *empirically robust*? This constitutes the second challenge for the Relational Model and it will be addressed in chapters 5 and 6.

Chapter 5 focuses on the cognitive feasibility and the practical importance of *relational* mindreading. First, relational mindreading will be incorporated into current explanatory theories that explicitly target the subpersonal implementation basis of our folk psychological competence. Second, powerful considerations will be presented as to why relational mindreading should be considered the psychological linchpin of human discursive practice. Relational mindreading plays a vital role in the quick and reliable attribution of propositional attitudes in quotidian, holistically structured contexts of interpretation, it easily accounts for the attribution of knowledge implied by ordinary explanations of action, and it guides children's first attempts at discursively interacting with others. In all these respects, the Relational Model gives a much more plausible picture of the psychology of common sense-making than the Belief-Desire Model.

In chapter 6, attention shifts towards the important complementary functions of *representational* mindreading. As indicated in the previous section, representational belief-desire psychology enables us to make sense of others even when their attitudes fail to align with our common assessment of the world. This makes it rather superfluous at the ground level of human discursive engagement, and rather impractical when used for predictive and explanatory purposes from a strictly third-person point of view. The private conception of mind is of great value, however, for the management of our discursive engagements in difficult or problematic social situations, and it proves necessary for critical evaluation of the norms that shape those interactions.

The distinction between relational and representational mindreading, between the attribution of public and private mental states, cuts across our folk psychological practices. Why has it generally been ignored in philosophical practice? Chapter 7 returns to the reflective fallacy. Having established both the conceptual validity and the empirical robustness of relational mindreading, it can now be fully appreciated why the in principle availability of belief-desire explanations of intentional action should not be explained in terms of conceptual entailment or psychological requirement. The account presented

in chapter 6 moreover suggests that it be understood as a particular manifestation of the evaluative function of belief-desire psychology in human social practice. The in principle availability of belief-desire explanations of intentional action can be explained in terms of the social dynamics of philosophical practice itself.

2

Goal-Reason Psychology

2.1 Introduction

One of the primary jobs of philosophy is to provide an accurate characterization of the phenomena we want to explain. In the debate on folk psychology, however, most philosophers seem to have been more interested in providing explanations than in directing their conceptual scrutiny to the *explanandum* their theories were targeting. In a rush to enter into the debate, many started from the allegedly commonsensical assumption that folk psychology is belief-desire psychology. But this is itself a substantial theoretical claim, presenting us with a model *of* folk psychology that is certainly debatable as such.

In this chapter I start from what I consider to be a more neutral characterization of folk psychology, i.e. that, amongst other things, it conceptualizes behavior in terms of goals and reasons for action. The question then becomes whether such goal-reason psychology is best understood as belief-desire psychology. The preliminary conclusion at the end of this chapter will be that it is not. Commonsense goal-reason psychology is first and foremost goal-reason psychology, and wielding it consists in drawing connections between agents, their goals and their reasons.

The structure of this chapter is as follows. First, I will discuss the Belief-

Desire Model of action explanation in action theory. Then, in the third section, I will show how this account has been adopted almost unanimously as an account of the psychology of folk psychology, i.e. of the attribution of goals and reasons in everyday social practice. The fourth section explores the relation between goal-reason attribution and belief-desire attribution and concludes with the suggestion that the former does *not* require the latter. This carries over to section 5, where it is proposed that the relation between goal-reason psychology and belief-desire psychology should not be explained psychologically (in terms of presupposition) or logically (in terms of entailment), but rather *socially* (in terms of the rationale of belief-desire ascriptions in social practice). This will give us a first insight into the nature of the reflective fallacy introduced in the previous chapter. Section 5 thereby also gives a first impression of the Relational Model of folk psychology as an alternative to the Belief-Desire Model. Section 6 coins the term 'relational mindreading' as a *non*-projectionist philosophical characterization of commonsense goal-reason attribution and places it in the context of the mindreading literature. This then sets the stage for the chapters to follow.

2.2 The Belief-Desire Model of Action Explanation

In order to make sense of an intentional action of another person, it is often important that we understand the purpose or goal of the action. It may also be important that we understand the reasons for which the action is performed. Thus, if John is walking down the street, it may be relevant that we find out that e.g., he is walking towards the supermarket in order to buy some milk. And it may be equally relevant that we know that he intends to do so because he has run out of milk, or because he won't be able to go to the supermarket tomorrow, or because he's having some friends over for lunch (his reason for having this goal). We can also think of situations in which it is important that we find out about the reason for his *walking* towards the supermarket (rather than e.g. taking his car) or his walking on *this particular street* (rather than taking the usual route). Perhaps his car is at the garage, so that walking is the best means of transportation given the circumstances (John's reason for walking in order to achieve his goal). Or maybe construction work is blocking his usual route, so that taking this route is the shortest way to the supermarket at the moment (his reason for walking on this particular street in order to achieve his goal).

In the example above, John's goal is given by a description of the intended

result of his action: an event (in order to *buy some milk at the supermarket*). His reasons are presented as states of affairs (because *he has run out of milk*) or facts (that *walking is the best means of transportation given the circumstances*). Goals and reasons thusly characterized are the things we often mention by using 'in order to', 'because' or 'for the reason that' clauses when giving answers to questions why regarding our actions. In what follows I shall use the terms 'goal' and 'reason' accordingly. Thus, goals are the things we intend to bring about by acting: events, or states of affairs. Reasons are the things that we consider to obtain (have obtained, will obtain) and to favor adopting certain goals and performing certain actions as a means to achieve them: what we perceive to be states of affairs, events or facts that solicit a response.[1]

Taking our commonsense explanations of action at face value, it seems we often engage in what we might term 'goal-reason psychology': we interpret each other's actions in terms of the goals we intend to achieve in the light reasons that make these goals and their means worth accomplishing. Even if we do not know the specific goal of another person's action, there often seems to be an implicit assumption to the effect that the action has some goal or other. And knowing the goal of an action often seems to carry the assumption that the goal is adopted and the particular action performed for certain reasons. Thus, seeing John walking down the street, we'd probably assume that he is going somewhere to do something for some reason, and that his walking down this street is somehow conducive to his reaching his goal. It is assumptions of this kind that motivate us to ask him, on certain occasions, where he is going, what he is going to do there and why, etc. From an interpretative point of view, in short, many intentional actions imply goals and reasons for which they are performed.

One of the central themes in contemporary philosophy of action is to obtain a precise understanding of these implications. According to what has become the standard account, goal-reason explanations of action evolve around the concepts of belief and desire. Proper understanding of some behavioral event as a genuine instance of intentional action, i.e. an action directed at

1 For similar use of the notion of a reason for action see e.g. Dancy (2000), Bittner (2001), Stoutland (2007), Alvarez (2010). There is a longstanding tradition in the philosophy of action according to which an agent's 'motivating' reasons for action are psychological states of the agent: belief-desire (pro attitude) pairs that *represent* the things to be achieved and the things taken into consideration or responded to by performing the action (cf. Davidson 1963/2001a, Smith 1987, 1994). For now, this can be regarded as a mere terminological difference. The issue that concerns us here is whether the attribution of goals and reasons, in my sense of the term, *implies the attribution of* such belief-desire (pro attitude) pairs (see below). See the appendix to this chapter for further discussion.

achieving certain goals in the light of certain reasons, demands that we regard it as being informed by appropriately structured beliefs and desires: beliefs representing the agent's reasons and desires representing the agent's goals.

Davidson once claimed that "In order to understand how a reason of any kind rationalizes an action it is necessary [...] that we see, at least in essential outline, how to construct a primary reason." (1963/2001a p. 4)[2] Davidson introduces the technical notion of a primary reason – not to be confused with the notion of a reason for action introduced above – as consisting of a pro attitude of the agent towards actions of a certain kind and a belief of the agent that his action is of that kind.[3] Thus, in the example above, John's primary reason could consist of his desire to perform a certain type of action, namely getting some milk, and his belief that his walking down the street satisfies this description of this action type, that is: that his walking down the street is a way of getting some milk.

Davidson's claim reads that in order to understand an action as an intentional action performed for a reason, it is necessary that we take there to be some pro attitude/belief pair or other that rationalizes the action. In many cases, finding out the specific pro attitude/belief pair is highly relevant, and explicit 'construction' of such primary reason may be required. But even if we do not know the specific pro attitude-belief pair that informed the agent's action, conceiving of her performance as an intentional action demands that we assume, 'at least in essential outline', some such pair to appropriately describe the event. Thus, in order to understand John's walking down the street as an intentional action, it is necessary, according to Davidson, that we see it as being informed by some pro attitude-belief pair or other.

More recently, Michael Smith (1987; 1994; 1998/2004) has defended what he terms the 'Humean' account of action explanation. A Humean explanation of an intentional action proceeds by citing a suitably structured belief-desire pair, consisting of a desire toward some end and a belief regarding the means.[4] In John's case above: his desire to go to the supermarket and his belief that walking down the street is a means to achieve that end, or his desire to buy some milk and his belief that going to the supermarket is a means to achieve

2 The quote reads 'and sufficient' between the brackets. This sufficiency claim is not relevant for our present purposes, nor is it plausible, as Davidson himself later recognized (e.g. 1978/2001a).

3 "R is a primary reason why an agent performed the action A under the description d only if R consists of a pro attitude of the agent towards actions with a certain property, and a belief of the agent that A, under the description d, has that property." (Davidson, 1963/2001a, p. 5)

4 Cf. Smith (1987, p. 36): "R at t constitutes a motivating reason of agent A to φ iff there is some ψ such that R at t consists of a desire of A to ψ and a belief that were he to φ he would ψ."

that end, etc.

Smith's Humean account makes two claims. The first is "that it is always possible to construct a Humean, belief/desire explanation of action." (1998/2004, p. 155) The second claim is that "once we see the central place occupied by Humean belief/desire explanations, we see that all the other explanations we give simply supplement this basic Humean story." (p. 156) The first claim is rather non-committal; it merely states that folk psychologically competent interpreters have the capacity to parse an agent's goals and reasons in terms of her beliefs and desires when the social situation demands it.

It is the second claim that makes the Humean account more substantive. For it says that Humean belief-desire explanations form the *explanatory core* of all explanations of intentional action we provide in daily social life, including explanations in terms of the agent's goals and reasons, in the perfectly ordinary sense identified above. Thus, when we explain an action by describing a state of affairs, e.g. John's going to the supermarket because he's run out of milk, we do not, Smith argues, "compete with the basic Humean story in terms of desire and belief, but rather *presuppose* and *add to it*." (p. 157, emphasis added) He later changes this into the more careful claim that "commonsense explanations [of action] presuppose the *availability* of a standard, Humean, belief/desire explanation." (p. 176, emphasis added) Thus we seem to arrive at the same core claim that Davidson made, be it in terms of desires rather than the general class of pro attitudes.[5] In order to understand how a reason of any kind explains an action it is necessary that we see (we have to presuppose), at least in essential outline (the availability of), how to construct a primary reason (a Humean belief-desire explanation). This is what I shall term the 'Belief-Desire Model' or 'BD-Model' of action explanation.

This section started with the assumption that making sense of an intentional action often demands that we engage in 'goal-reason psychology': that we interpret the action as having a goal and being performed for reasons. The BD-Model takes this assumption one step further by stating that understanding an agent as having goals and acting for reasons in turn demands that we see him as instantiating appropriately interlocking beliefs and desires, or at least rests on appreciation of the fact that explanation in terms of such beliefs

5 A desire is just one kind of pro attitude, as Davidson used the latter term: "under [pro attitudes toward actions of a certain kind] are to be included desires, wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values…" (1963/2001a, p.4) The distinction between desires and pro attitudes is not very important for our purposes. In what follows, I will use the term 'desire' to refer to a broader class of pro attitudes, viz. those pro attitudes that have representational, propositional content (see section 2.4).

and desires is available in principle. On this picture, our commonsense understanding of an agent's action on the one hand, and his goals and reasons on the other, is brought together in virtue of their sharing a common presupposition: that the action is informed by the agent's beliefs and desires.

2.3 Folk Psychology as Belief-Desire Psychology

The BD-Model of action explanation in the philosophy of action has had a profound influence on philosophical treatment of folk psychology. In the most neutral sense of the term, folk psychology comprises the large body of commonsense explications, explanations, predictions, etc. we, 'the folk', generate in the service of making sense of each other in everyday social life. Philosophical interest in folk psychology ranges from more metaphysical concerns to more epistemological ones.

On the metaphysical side, the focus lies on the status of folk psychology in relation to (mature) cognitive neuroscience. On David Lewis's (1972) influential analysis, folk psychology is to be regarded as an implicit functionalist theory that specifies how mental states are causally related to sensory stimuli, motor output and other mental states. Mental state terms, as 'the folk' understands them, are implicitly defined by this commonsense functionalist theory. On this analysis, our folk psychological concepts are theoretical concepts indicating functional entities that occupy causal roles in the production of thought and action. Lewis's conceptual analysis of folk psychology as a term-introducing theory gave rise to heated discussions about the relation between this presumed folk psychological theory on the one hand and our scientific theories on the other, that is: those scientific theories that aim at explaining the production of thought and action. Some proposed a mesh between the two and argued for a vindication of folk psychology by cognitive science (e.g. Fodor 1987, Dretske 1988), others envisaged a clash that would result in the scientific elimination of folk psychology (e.g. Churchland 1981, Stich 1983), while still others suggested a slightly more relaxed interpretation of Lewis's analysis and proposed a rather independent and peaceful co-existence of folk psychology and cognitive neuroscience (e.g. Dennett 1987, Jackson and Pettit 1988). A minority took a different approach altogether by taking issue with the heart of Lewis's proposal. They challenged his construal of folk psychology as a theory and thereby tried to disarm the heated debate about the fate and

future of folk psychology (e.g. Wilkes 1991, Baker 1995).[6]

Without exception and regardless of their differences, all these accounts are explicitly framed in terms of the attitudes of belief and desire. At first sight, this may seem a rather innocent consequence of their preferred choice of subject: the ontology of these propositional attitudes. Are the folk concepts of belief and desire functionalist concepts, do they purport to say anything about what's going on inside our heads, and if so, how does that fare with what our best scientific theories have to say about that? These are surely important issues in the philosophy of mind that justify specific interest in our belief and desire concepts. But closer inspection reveals that this focus was also motivated by an assumption about the *structure* of folk psychology, viz. that it is, at its core, a belief-desire psychology. Thus, Baker explicitly states that "Although commonsense psychology encompasses much more than propositional attitudes […] belief-desire reasoning forms the core of commonsense psychology." (1999, p. 3) In similar fashion, Fodor tells us that "the theory from which we get [our] extraordinary predictive power is just good old commonsense belief/desire psychology" (1987, p.3), Churchland targets propositional attitudes such as belief and desire as "the principal elements of common-sense psychology" (1981, p. 67), and Dennett holds that in exercising our folk psychological capacities, "we approach each other as intentional systems, that is: entities whose behavior can be predicted by the method of attributing beliefs, desires and rational acumen." (1987, p. 49)[7] Underlying their interest in the ontology of the propositional attitudes, then, we find a commitment relating to the *psychology* of folk psychology, viz. that the ascription of beliefs and desires to one another is the central engine of our practice of explicating, explaining and predicting each other's thoughts and actions in daily social life.

The psychology of folk psychology is the target of the (descriptive) epistemological debate on folk psychology (cf. Goldman 1993; 2000). Here the central question concerns our adult capacity to generate and understand folk psychological, explications, explanations, predictions, etc. What precisely does it consist in? Traditionally, the debate was defined by two main

6 Some writings of Dennett seem to point in this direction as well. See especially his distinction between folk psychology as 'craft' and as 'ideology' (Dennett 1991).

7 Dennett is rather explicit about the actual interpretation activity in folk psychological practice: "First you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many – but not all – instances yield a decision about what the agent ought to do; that is what you predict the agent *will* do." (1987, p. 17)

positions.[8] According to the first, our explicit attempts to make sense of other people are effectuated by implicit deployment of a 'theory of mind' (Premack and Woodruff 1978). This general position, termed the 'Theory Theory' (TT) (Morton 1980), is the epistemological counterpart of the Lewisian analysis of folk psychology.[9] Theory theorists argue that our understanding of other people's actions requires making use, in some way or other, of a theory that specifies the functional roles of beliefs, desires and other psychological states. We use this folk psychological theory to make inferences to the best explanation as to what an agent's mental states are and how they give rise to behavior. It is not a theory that we have conscious access to or that we can state explicitly on demand. Rather the theory is said to be tacit or implicit, meaning that it is stored in our brains, subconsciously guiding the information processes that lead up to our explicit judgments about the actions and mental states of others. (cf. Davies 1994, Stone and Davies 2001, Ravenscroft 2003)

This theoretical approach to the epistemology of folk psychology was the default position in philosophy until the mid 1980's, when Jane Heal (1986) and Robert Gordon (1986), later followed by Goldman (1989), proposed a 'Simulation Theory' (ST) of folk psychology. On this proposal our capacity for explaining and predicting other people's thoughts and actions is not mediated by a theory that specifies the functional roles of mental states, but rather proceeds by using our own (practical) reasoning skills and emotional responses as a model of the minds of others. Thus, instead of theorizing about other people's reasoning, we simply 'replicate' (Heal) that reasoning ourselves, thus arriving at their conclusions by letting our own reasoning skills determine which are the proper and relevant inferences to make. Explaining or predicting another person's action, we identify with them (Gordon) or place ourselves in their shoes (Goldman) and let our own practical reasoning skills and other response mechanisms work on the pretend context of action in order to find out why the agent performed the action or what her next move is going to be.

8 Goldman (1989, 2006) distinguishes a third position he terms 'rationality theory', which he attributes to Davidson and Dennett. Rationality theory has never played an important role in the epistemological debate on folk psychology, mainly because its two leading proponents did not directly participate in it. In fact, their concerns were mainly with the metaphysical aspects of folk psychology. Like Lewis, they put forward an account of folk psychological concepts in order to make progress on the metaphysics of mind. The leading idea is that the use of folk psychological concepts is governed by norms of rationality that make folk psychology indispensible but in principle unfit for description, or even approximation, of sub-personal processes (Dennett) or nomological regularities of physics (Davidson). Yet, they also appear to be making claims about the actual folk psychological activity of interpreting other people, Dennett in particular (see n. 7).

9 In fact, one could argue that it is the epistemological *implication* of the Lewisian analysis. For an interesting discussion on this score, see Jackson, Mason and Stich (2009).

Exploiting our folk psychological capacity for making sense of other people does not primarily consist in the deployment of tacit knowledge of a functionalist theory of mind. Our most important interpretation skills are 'process driven' rather than 'knowledge driven' (Goldman 1989).

Recently, a third general approach has been introduced in the debate. This approach stresses the interpretationist insight (e.g. Davidson 1970/2001a, Dennett 1987) that folk psychological practice is a *normative* practice. Folk psychological practice is not only a matter of explanation and prediction, but also, and most importantly, of justification, regulation and education (cf. Morton 2003, McGeer 2007, Gallagher and Hutto 2007, Hutto 2008a, Zawidzki 2008, Andrews 2009). According to Hutto (2008a), for example, folk psychological practice is an essentially *narrative* practice. He attacks what seems to be a implicit assumption behind many of the accounts discussed so far, viz. that the core business of folk psychology is the prediction and explanation of action from a spectatorial, third-person standpoint. According to Hutto, we learn to participate in folk psychological practice in participatory, second-person contexts and these contexts remain the primary *raison d'être* of adult's deployment of their folk psychological skills. The idea behind this is that these skills are brought to bear only in the sorts of cases in which we are surprised or perplexed about the actions of other people; normally, our culturally based expectations obviate the need to explicate their ways. In instances of surprise or perplexity, second-person interaction serves to normalize a seemingly abnormal course of action by placing it in the context of a larger personal narrative. (Hutto 2008a, p. 32-40). The story the agent tells is suited to her particular situation, it does not necessarily generalize to other people or situations. It functions to normalize her actions in the face of her apparent failure to meet cultural norms or the interpreter's expectations. It is their particularity, normativity and primary second-person use that renders folk psychological explications rather non-theoretical, and better characterized as instances of story telling.

This panoply of philosophical approaches to the psychology of folk psychology is impressive, and some of them contain important insights that will be developed in later chapters of this book. For now, however, we should return where we left off in our discussion of the metaphysical debate on folk psychology. There it was noticed that concerns regarding the status of folk psychology were accompanied by a claim about its structure: that it is, at its core, a belief-desire psychology. This claim, in turn, seemed to be inspired by an assumption regarding the psychological requirements for generating and understanding folk psychological explanations and predictions: that it demands the ascription of beliefs and desires. The same line of thought has

also been driving the epistemological debate on folk psychology. Thus Davies and Stone (1995, p. 2), for example, state that the conceptual repertoire that constitutes folk psychology "includes, predominantly, the concepts of belief, desire and their kin [...] the propositional attitudes," and Hutto (2008a, p. 3) targets folk psychology 'stricto sensu' with explicit focus on belief-desire psychology, agreeing with Baker's claim quoted above that "belief-desire reasoning forms the core of commonsense psychology." (Ch. 1, n. 5)

These assumptions regarding the core structure of folk psychology are taken further by others, transformed into the stronger claim that belief-desire ascription is a psychological requirement for generating explanations and predictions of the actions of others. Nichols and Stich (2003, p. 4), for example, say that "the central concepts implicated in mindreading' are 'belief, desire and intention." These claims are taken as a starting point for more elaborate hypotheses about the psychological mechanisms underlying our folk psychological capacities. On Goldman's version of ST (2006, pp. 44-45, 185-188), action explanation and prediction requires the classification of pretend beliefs, desires and decisions as such. And Fodor (1992, p. 283) speculates that normal socio-cognitive development "eventuates in the child's internalization of a tacit "metacognitive" intentional psychology: specifically, in the internalization of some version of the folk psychological theory that an agent's behavior is normally caused by his beliefs and desires."[10]

Fodor's theory is a version of the 'modular' theory that has played an important role within developmental psychology. On this story, children have a (partly) innate 'theory of mind mechanism' that deploys the concepts of belief, desire and some other propositional attitudes and predisposes the normally developing child to pay selective attention to mental states of others, thereby enabling them to attribute such states in the course of action explanation and prediction (e.g. Baron Cohen 1995, Leslie 1994, Scholl and Leslie 1999). Other psychologists have put forward an radically empirist account that portrayed children as 'little scientists', advancing, testing and rejecting increasingly accurate theories about the behavior of others in relation to their environment. In their first few years of life, children come to reject relatively simple desire-based and belief-based theories in light of overwhelming countervailing evidence, eventually arriving at a full-blown belief-desire theory that form the basis of their social functioning throughout the rest of their lives.

10 Consider also the numerous 'boxologies' of third-person mindreading in Stich and Nichols (1992, 1995, 1997) and Nichols and Stich (2003, chapter 3), or Currie and Sterelny (2000, pp. 145-146), who maintain that "our basic grip on the social world depends on our being able to see our fellows as motivated by beliefs and desires we sometimes share with them and sometimes do not."

(e.g. Gopnik 1996, Gopnik and Wellman 1992, Gopnik and Wellman 1994, Gopnik and Meltzoff 1997). Although these two general approaches differ in their account of the developmental trajectory, they too share the same core assumption regarding the end-stage of this trajectory: that our adult folk psychological capacity to interpret other people's actions is essentially a matter of attributing beliefs and desires.[11]

In general, then, there has been a strong tendency, both in the more philosophical discussions between theory theorists, simulation theorists and more recent contenders, and in the more psychological discussions between nativists and empirists, to characterize our adult folk psychological capacity for generating and understanding action explanations as centering around the capacity for belief-desire attribution. Accordingly, explanations in terms of goals and reasons as characterized above – things to be achieved in the light of things that make them accomplishable and worth accomplishing – have to be regarded as essentially elliptical expressions of proper belief-desire explanations. Interpreting another person's action in terms of his goal and reason aims at laying bare the beliefs and desires that inform the action. Theory theorists claim that this process is guided by lawlike psychological generalizations, chief amongst which is the 'central action principle': "if A wants P and believes that doing q will bring about p, then ceteris paribus, A will q." (Borg 2007, p.6)[12] Whereas simulation theorists oppose to the idea that the interpreter needs to have (tacit) knowledge of such action principles (but see Ravenscroft 2003), most of them seem to agree that the simulation procedure needs to at least *mirror* these principles: no action explanation or prediction without offline processing, classification and attribution of (pretend-) beliefs and desires. And although Hutto (2008a) sharply distances himself from a cognitivist rendering of the interpretation processes involved, he does regard "the way beliefs and

11 Gopnik and Meltzoff (1997, p. 126), for example, claim that our mature theory of mind, "(...) has many complexities but also a few basic causal tenets (...). These tenets are perhaps best summarized by the "practical syllogism": if a psychological agent wants event y and believes that action x will cause event y, he will do x." On the nativist side, consider Scholl and Leslie (1999, p. 132): "A theory of mind refers to the capacity to interpret, predict, and explain the behaviour of others in terms of their underlying mental states. It is an ability that all normal humans enjoy, and seems to manifest itself in early childhood. This capacity is inherently 'metarepresentational', in that it requires one not only to employ propositional attitudes, but to employ them *about* propositional attitudes, for example having beliefs about (others') beliefs." Cf. "in everyday life we make sense of each other's behaviour by appeal to a belief-desire psychology" (Frith and Happé 1999, p. 2); "a prediction of behavior requires additional ascription of desire, the integration of belief and desire, and the inferring of a resulting action." (Leslie, German and Polizzi 2005, p. 50)

12 Cf. Botterill (1996, p. 115): "If belief-desire psychology has a central principle, it must link belief, desire and behavior. It could be formulated like this: [action principle] An agent will act in such a way as to satisfy, or at least to increase the likelihood of satisfaction, of his/her current strongest desire in light of her beliefs." See also the quote of Gopnik and Meltzoff in n. 11.

desires conspire to motivate actions' as comprising 'what we might think of as the "core principles" of intentional psychology," structuring the folk psychological narratives we tell each other (p. 29).

Within the debate on folk psychology, commonsense goal-reason psychology has been characterized as a belief-desire psychology through-and-through. On the assumption that the attribution of goals and reasons hinges on the capacity to attribute beliefs and desires, it is our proficiency in belief-desire psychology that has been targeted as the *real* explanandum.[13]

2.4 Relating People to Their Goals and Reasons

Let us retrace our steps to where we began: with the rather innocent assumption that when we make sense of other people's actions we often assume there to be goals and reasons for which these actions are performed.

A practical reason is something that favors an action and may render it the appropriate or right thing to do. Thus his having run out of milk may be a reason for John to buy some milk. And the fact that he can buy milk at the supermarket may be a reason for him to go to the supermarket in order to buy some milk. Reason giving explanations therefore have a so-called 'word-to-world' direction of fit (e.g. Searle 1983). A reason explanation given in response to a question why should adequately describe some feature of the world that made the action appropriate or right: the explanation should contain a description that 'fits' a favorable condition in the world. A goal, on the other hand, is something that determines the success conditions of an action. If John is walking towards the supermarket in order to buy some milk, then the success of his walking towards the supermarket can be measured by the extent to which it will enable him to buy some milk there. A goal has a 'world-to-word' direction of fit: an explanation in terms of one's goal given in response to a question why determines how the world is to be changed as a result of successfully carrying out the action.

In order to understand John's action in terms of his goal and his reasons, we need to see John inter alia as *responding* to his having run out of milk by *intending* to go to the supermarket to buy some milk. Understanding John as responding to a reason by intending to carry out and complete an action is a

13 There are a few notable exceptions, e.g. Gordon (1987, 2000, 2001), Perner (e.g. 1991), Perner and Roessler (2010) and Ratcliffe (2006, 2007, 2009). See chapter 4 for discussion of these accounts.

typical example of 'intentionalistic' interpretation: it requires that we perceive John as being *intentionally directed at* his reason and his goal. John's goal is something we may regard as an event bound to happen or likely to occur in the near future: his buying some milk at the supermarket. On the assumption that John has actually run out of milk, we may conceive of John's reason is the state of affairs of his having run out of milk. Interpreting John's action as an intentional action requires that we see him as responding intentionally to some state of affairs by intending to bring about some event. It thus involves drawing two kinds of 'intentional connection' between the state of affairs constituting his reason, the event constituting his occurrent behavior and the event constituting his goal: an intentional connection with a 'mind-to-world' direction of fit between his action and his reason, and an intentional connection with a 'world-to-mind' direction of fit between his behavior and his goal. The important issue is whether drawing these intentional connections amounts to belief-desire ascription.

There are three features of belief-desire ascription that are relevant for our present purposes: it results in the attribution of (i) representational mental states with (ii) propositional content that (iii) have a particular direction of fit. Starting with (iii), beliefs are said to have a mind-to-world direction of fit, desires a world-to-mind direction of fit (e.g. Searle 1983). A true belief answers to the way the world is, it fits with the world. A false belief should be discarded and changed to fit with the world, and not vice versa. By contrast, a desire can be realized, and if so, the world fits with the desire. An unrealized desire should not be discarded simply because it is not realized. Rather, a desire may provide ample reason to change the world to fit with it, but not vice versa. This seems to fit nicely with the conclusion reached above about an agent's reasons and goals. Beliefs have the same direction of fit as the intentional attitude of taking into consideration or responding to a reason and desires have the same direction of fit as intending to achieve a goal.

Turning to (ii), beliefs and desires are often referred to as propositional attitudes. This means that in ascribing beliefs and desires, the contents of the attributed mental states should be propositionally articulated, i.e. have the appropriate structure to be truth evaluable and to stand in inferential relations to the contents of other mental states of the agent and the meaning of her linguistic utterances.[14] Interpreting an agent's action in terms of her goals and

14 There is some discussion whether desires always have propositional contents. Some hold that desires may be directed at intentional objects rather than situations or states of affairs (e.g. 'I desire that piece of chocolate' rather than 'I desire that I have that piece of chocolate'). But this treatment seems strained in the case of belief. Beliefs are true or false in virtue of their propositional

reasons also requires sensitivity to the propositional structure of her intentional attitudes. It requires that we are able to place the action within the scheme of practical reasoning and within the discursive practice of giving and asking for reasons. Even if the action did not result from conscious practical deliberation (which, arguably, is often the case), making sense of the action as a second or third person interpreter may involve some practical reasoning about the agent's goals and reasons. And knowing how to engage in reason discourse, ask the appropriate why-questions and understand the agent's responses as justifying because-answers, demands that we understand her reasons and her goals in truth-evaluable, propositionally articulated fashion.

Taken together, features (ii) and (iii) of belief-desire ascription mirror two important requirements of the attribution of reasons and goals. From this, I expect, many would want to conclude that identifying an agent's reasons and goals consists in determining, inter alia, the contents of some of her beliefs and desires.

But this is too quick. Arriving finally at feature (i) of belief-desire ascription, it should be realized that parsing the agent's mind into beliefs about his reasons and desires about his goals results in an articulation of the way the world is (was, will be, should become) according to the agent in particular. Genuine ascription of beliefs, whether true or false, or desires, whether realistic or unrealistic (appropriate or inappropriate, etc.), demands that we keep a distinction between the ascribee's subjective view on the world and the world itself, as it reveals itself to us. Ascribing to the agent a belief about his reason, we differentiate between what the agent *believes* to be the worldly conditions that favor his action, and what the worldly conditions really are. When we ascribe a true belief in the course of explaining an action, we lock his view onto the real structure of the world, as we take it to be. Likewise, in attributing a desire for and a corresponding intention towards a certain outcome, we distinguish between the agent's representation of how the world is supposed to be changed as a result of his action, and the way we expect the world to be changed by that action. By ascribing a realistic desire, we project the contents of the agent's desire onto our own expectations regarding the future course of events.

What we need is an argument to the effect that maintaining a distinction

content; objects cannot be true or false. And in the case of desire, it is the ascription of the propositional attitude of desiring that is suited making sense of the goals people adopt in response to reasons and the practical reasoning that may be involved in such sense-making. It is in virtue of their propositional form that the contents of desires can enter into the appropriate logical relations with other propositional states that make up instances of practical reasoning. Cf. Hutto (2008a, p. 2)

between the agent's representations of the world and how the world presents itself to us is required for interpreting his action in terms of his goals and reasons. But it is far from obvious that there is such a requirement for goal and reason attribution *simpliciter.* In the case of John walking down the street, for example, we'd do perfectly fine, it seems, by simply drawing relations between a state of affairs (his having run out of milk), a present event (his walking down the street), and a future event (his buying some milk at the supermarket), given a fact (that he can buy some milk by going to the supermarket).

Here, the term 'relation' should be understood in the strict sense of implying the past, present or future existence or reality of both (all) *relata*. Of course, this implication should be understood from our point of view as interpreters of John's action. We would be relating John to what *we* take to be the states of affairs John is responding to and what *we* expect will be brought about by John's action. We could be mistaken about the things we interpret John as responding to or make the wrong 'predictions' about the future course of events that will follow as a result of John's action.[15] What matters is the *interpretative* distinction between understanding others *directly* in terms of the world (past, present or future) as it presents itself to us and understanding them *indirectly* in terms of their *representations of* the world (past, present or future) as it presents itself to us.

As argued above, the relations drawn between John's and the world around him would have to be *intentional* relations, of *responding* to his having run out of milk, of *being aware* that he can do so by buying some milk at the supermarket, of *intending* to buy some milk at the supermarket, etc. Crucially, however, these intentional states would *not* be attributed as representational

15 The term 'prediction' may be somewhat misleading, because our expectations about the future in relation to John's action need not be based on any *further* evidence than the verbal expression of his intention to buy some milk at the supermarket when we ask him what he is up to. Our expectations regarding the effectiveness of other people's actions are perhaps best understood in analogy to first-person avowals of intention, as in 'I shall buy some milk in the supermarket'. Anscombe (1957) suggested that such avowals are predictions in the sense that they provide descriptions of something to occur or obtain in the future, but not in the sense that they are based on some kind of (introspective or behavioral) evidence (see Hamilton (2008) for a treatment of avowals of intention along these lines). Accordingly, second- or third-person understanding of avowals of intention would take the form of attributing *practical commitments* to make it happen that something occurs or obtains (cf. Brandom 1994, ch. 4). On this account, expectations about the future based on the attribution of practical commitments (assuming that the agent has the reliable responsive disposition to act accordingly) are not the same in kind as predictions about the future based on ordinary empirical evidence (e.g. predicting that the vase will fall on the ground and break when pushed off the table). Yet such expectations about other agents' intentional behavior may generate the same degree and feeling of certainty as predictions about non-agential or non-intentional occurrences. In this sense, interpreting agents' behavior in terms of their goals can be characterized as *implying* the occurrence or presence of some future event or state of affairs to be brought about by their actions.

states. Understanding John as walking toward the supermarket in order to buy some milk because he's run out of milk, we would be making sense of his action in terms of *relational* mental states: psychological states attributed to John that relate him to his goal and his reasons.[16]

2.5 The Fallacy: A First Assessment

Let us now return to the two claims Smith makes in defense of the BD-Model of action explanation. The first says "that it is always possible to construct a Humean, belief/desire explanation of action." (1998/2004, p. 155) Accordingly, whenever an action is explained in terms of a goal at which it is directed and a reason for which it is performed, there is an explanation of that action in terms of, *inter alia*, a Humean belief-desire pair. Call this conditional C. The question is how we should explain the truth of C.

Smith chooses to explain it by arguing that the Humean explanation (or at least its availability) is *presupposed* by the explanation in terms of goals and reasons. This is his second claim. Accordingly, "once we see the central place occupied by Humean belief/desire explanations, we see that all the other explanations we give simply supplement this basic Humean story." (*ibid.*, p. 156) Thus, when someone explains John's going to the supermarket by saying that he does so in order to buy some milk because he's having some friends over for lunch, he not only presupposes the existence of John, his friends and the supermarket, but also John's desiring to buy some milk, his believing that he's having some friends over for lunch, that he can buy milk by going to the supermarket, etc. (or at least the presence of some such representational states of John).

As we have seen, the BD-Model has been adopted within the debate on folk psychology as the claim that in attributing a goal and a reason to another person, we must be (tacitly) ascribing Humean belief-desire pairs. Thus, the truth of C is explained in terms of *psychological requirement*: the attribution of goals and reasons requires the ascription of beliefs and desires. Adherents of the BD-Model often seem to be making an even stronger claim: that it is simply a matter of a priori conceptual truth that goal-reason explanations imply belief-desire explanations. The idea would be that the conceptual structure of

16 In what follows, I shall be using the term 'state' in a metaphysically rather noncommittal way. In particular, I do not want to imply that mental states are 'states' in a sense opposed to episodes.

our folk psychology is such that having goals and reasons for action *entails* instantiating desires representing these goals and beliefs representing these reasons.[17] According to this line of thought, the truth of C should thus be explained in terms of *conceptual entailment*.

But there is another option for explaining the truth of C. Rather than trying to explain it *psychologically* by saying that the antecedent presupposes the consequent or that the former explanatory strategy requires tacit deployment of the latter, or *logically* by claiming that the antecedent entails the consequent, we could explain it in terms of the *social* dynamics of the explanatory practice itself.

There is a rule in boxing that allows the coach of a fighter to throw in the towel at any time during the fight. Attorneys in American television series have the prerogative to shout 'Objection your honor!' during interrogation of a witness by the other party. The coach's and attorney's entitlement to make use of these procedures co-constitute the practices they are involved in, a boxing fight and a legal trial, respectively. And there are clear *rationales* for their entitlement to make use of these procedures. It is for the safety of his pupil that the coach is always allowed to throw in the towel. And, in general, it is for the sake of a fair trial that attorneys may object to the method of interrogation of a witness. Thus, there is a way of making sense of the fact that 'making a certain move' is always possible in a certain practice, by appeal to the rules and rationales of the practice itself.

Analogously, we could argue 1) that folk psychological interpreters can always make a move in the 'game of giving and asking for reasons' (cf. Sellars 1954/2007; Brandom 1994) from explanations in terms of the agent's goals and reasons to Humean belief-desire explanations and 2) that this fact is explained in terms of rationales revealing why this 'procedure' has become part of our social practice.

Consider a surgeon who, after having tried to reach an organ ventrally according to standard protocol, decides to take the dorsal route with a different surgical instrument, or a soccer coach who substitutes a defender for an attacker at the end of the match. Such changes in technique or strategy to solve a problem or win a game are common phenomena. But the surgeon's use of the first technique does not presuppose (let alone entail) the second tech-

17 To a certain degree, this seems to be inspired by the thought that motivating reasons are belief-desire pairs, rather than what is represented by certain beliefs (see appendix). This thought can only cause confusion in the present context. What matters is whether according to the folk psychological conception of intentional action, acting on reasons in my sense of the term, i.e. events, facts or states of affairs that solicit a response, entails having beliefs about those reasons.

nique, not even it's availability. Of course, the second technique *is* available – it belongs to her surgical repertoire and the instrument required for it lies on the table – but the surgeon need not appreciate this fact while she is using the first technique, not even tacitly. The availability of the second technique may simply first presents itself when the ventral route starts posing problems. The same goes for the soccer coach. The offensive strategy is always available, but it may come to mind only when the situation demands it, e.g. when the other team has scored a goal.

Analogously, we could regard the shift from an explanation in terms of the agent's goals and reasons to a Humean belief-desire explanation as a change in interpretation technique or explanatory tactics. It is a move that, as the game of giving and asking for reasons is set up, is always available for competent participants, but which skilled interpreters are prone to make only at certain crucial moments, viz. those moments in which social interaction starts becoming problematic in some way or other. In normal situations, however, the belief-desire strategy need not play any psychological role.

There is, of course, an important disanalogy between shifting surgical techniques or soccer strategies on the one hand, and changing to a Humean mode of interpretation on the other. The switch from making sense of someone in terms of her goals and reasons to interpretation in terms of her beliefs and desires is unique in its kind. It is a *representational* shift, a switch from attributing goals and reasons to attributing *representations* of goals (desires) and reasons (beliefs), from relating an agent *to* the world to relating to the agent's *view on* the world.

This brings us to the *rationale* of the switch to belief-desire psychology in ordinary discursive practice. The most important feature of belief-desire explanations is that they work irrespectively of the truth of the beliefs and the appropriateness or feasibility of the desires and resulting intentions of the agent. It is this neutrality regarding truth, appropriateness and feasibility that gives representational belief-desire psychology its primary *raison d'être*. When social interaction becomes problematic due to displays of apparently counter-normative or irrational behavior, it is often of crucial importance that we effectively manage our discursive engagements with one another and maintain the social balance that is required for successful coordination of our actions. As I shall explain in chapter 6, belief-desire psychology is a useful interpretative tool in such situations, precisely because it allows us to go beyond our common assessment of the world and to adopt a relatively disengaged attitude towards one another. Belief-desire psychology is our most sophisticated technique for dealing with cognitive conflict and conative divergence between

members within our community; it allows us to make our default explanations more precise, better suited to the particularity of individual agents.

This section started with the question how to explain conditional C: whenever an action is explained in terms of a goal at which it is directed and a reason for which it is performed, there is an explanation of that action in terms of, *inter alia*, a Humean belief-desire pair. I have made a first, sketchy attempt of how to answer this question *without* invoking the presupposition or entailment of the latter by the former. According to this alternative, belied-desire psychology is not the conceptual core of or the psychological engine *behind* our discursive engagements with one another, but rather a *complement to* that practice, a *secondary* interpretation technique specifically designed to deal with socially problematic situations.

This brings us to the title of this section. As indicated in chapter 1, the reflective fallacy is the fallacy of projecting a typically philosophical, reflective understanding of intentional action in terms of belief and desire onto our commonsense understanding of each other's intentional actions in terms of goals and reasons. Now if the present account is correct, belief-desire explanations can serve the social purpose *precisely of* taking a few steps back in order to reflect on each other's actions and evaluate them in a special way. Adopting this reflective mode of understanding for detached philosophical reflection is in line with its primary social function.

If the philosophy of action is indeed concerned, as Smith thinks, with the "attempt to state a principle that allows us to unify diverse [commonsense] explanations" (1998/2004, p. 155), then a Humean shift in interpretation is only to be expected: it simply reflects the rationale of belief-desire psychology in ordinary social practice that explains conditional C. Smith's Humean account runs the risk of committing the reflective fallacy when it regards this unifying 'Humean principle' as a *logical* principle underlying our commonsense concepts of having a goal and acting for a reason, or a *psychological* principle guiding our interpretative abilities, rather than a *social* principle that describes the dynamics of our discursive practices.

It may be considered a datum that people sometimes err, that they sometimes have very peculiar desires and that their aspirations are sometimes far from realistic. The BD-Model incorporates explanations of such aberrant cases into the explanatory scheme of default, veridical, appropriate and realistic cases. It thereby presents us with a unifying story of action explanation, but only at the expense of making our explanations of normal, i.e. well-informed, appropriate and realistic, actions subject to the requirements for explaining abnormal ones. This is symptomatic of the reflective fallacy. In the attempt

to articulate a concept that is at home in unreflective social practice, the philosopher seeks to use it for himself from a particularly reflective stance, thereby inadvertently altering the focus of investigation. It is a non-starter to test the adequacy of our default, unreflective conception of intentional action by putting it to use in what, from a folk psychological point of view, is an exceptional, detached and reflective mode of understanding.

2.6 Mindreading

The claim of this book is that the psychological basis of human discursive practice is relational rather than representational. On the Relational Model of folk psychology, we tend to interpret other people by drawing relations between them and the things in the world that constitute their goals and reasons. This is what I call 'relational mindreading'. It is *relational* mindreading because it consists in the attribution of relational mental states rather than representational mental states; it is relational *mind*reading as opposed to mere behavior or body reading; and it is relational mind*reading* because it requires sensitivity to the inferentially articulated propositional contents of the mental states attributed. The form of social understanding that results from relational mindreading has generally been neglected in the debate on social cognition. The conception of an agent being 1) genuinely intentionally directed at the world, as opposed to being merely behaviorally disposed to react to it, in 2) a propositionally articulated fashion, as opposed to more primitive, non-conceptual modes of responsiveness, without, however, 3) subjectively representing the world as being a certain way – this conception seems to have been overlooked entirely. Yet, as I shall argue, it is the conception that lies at the basis of our appreciation of each other as rational, discursive beings, beings who think, talk and act in response to reasons.

The reflective fallacy is at least partly to blame for this neglect. As theorists, we are so used to talking in terms of beliefs and desires and to focusing on their representational character, that the intelligibility of a relational notion of propositional attitude has simply escaped notice. Theorizing about reason attribution and acknowledging its propositional character, participants automatically switch to a representationalist belief-desire framework. But they do so without any argument, which gives the impression that there is an assumption at play to the effect that the propositional nature of content implies its representational nature. As I shall argue, however, there is no reason why this should be true from a commonsense point of view.

Another factor that has played a role in the neglect of relational mindreading is the ambiguity that has crept into our use of the term 'representation' in the debate. Sometimes it is used to indicate the subjective character of states like beliefs, that is: to highlight the fact that a belief captures the way the world is represented to the particular individual to whom it is ascribed. Attributing a belief to another person, an interpreter regards him as a subjective world-representer, with a particular view on the world that may or may not jibe with the world itself. At other times, the notion of representation is used merely as a functional notion, to indicate the causal or explanatory relevance of some internal state in the bringing about of certain behavior. Conflation of these two ways of using the notion of representation makes it difficult to appreciate the idea of relational mindreading (see chapter 3 and 4).

Relational mindreading should first and foremost be distinguished from representational mindreading. Representational mindreading, as I use the term, is just good old belief-desire psychology. It does exist, but its social function is not what many philosophers and psychologists have assumed it to be, i.e., to *ground* the practice of giving and asking for reasons. Belief-desire psychology is not the silent engine *behind* our rational social practices, but rather an adjunct *to,* enabling us to manage and to reflect *on* those practices (see chapter 6).

Relational mindreading should also be distinguished from what is sometimes referred to as 'theory of behavior' (cf. Povinelli and Vonk 2003, 2004; Perner and Ruffman 2005, Perner 2010). On this account, which has been offered as a possible explanation of the social behavior of primates and infants, the behavior of others is interpreted in terms of behavioral rules, rules stating behavioral dispositions in relation to environmental conditions without reference to mental states that mediate between stimulus and response. The states attributed by the relational mindreader are not mere behavioral dispositions. Nor do the features of the environment making up one end of the relation figure as *just that*, features of the environment; rather they feature as the contents of the states attributed, on the other side of the relation. These contents are moreover propositionally articulated, so that attributing a relational mental state allows the interpreter to draw inferences about the agent's sayings and doings that exceed any strictly behaviorist analysis.

Finally, relational mindreading should be distinguished from a variety of less sophisticated forms of triadic social responsiveness that do not exhibit sensitivity to the propositional structure of other people's intentional directedness. Here we can think of the ontogenetic precursors of sophisticated mindreading, such as perceiving other people's gaze as being object-directed

(Woodward 2003, Johnson, Ok and Luo 2007), assessing others' visual access to objects (Luo and Baillargeon 2007, Luo and Beck 2010), following pointing gestures (Woodward 2005), anticipating other agents' proximal goal-directed actions (Gergely and Csibra 2003), engaging in short cycles of object-directed joint attention (Tomasello et al. 2005), or responding to ostensive communicate signals and building generalizable referential expectations in communicative contexts (Csibra and Gergely 2007). It has been convincingly argued that these and numerous other relatively primitive forms of social responsiveness persist as the backbone of our adult socio-cognitive skills, supporting and targeting sophisticated mindreading when and where we engage in it (e.g. Gallagher 2001, Gallese 2007, Ratcliffe 2007, Gallagher and Zahavi 2008, Hutto 2008a, Bermudez 2009, Apperly 2011, Zawidzki forthcoming).

Many theorists these days distinguish between so-called 'low-level' and 'high-level' forms of social cognition.[18] The low-high dimension can be applied at least along two axes: 1) the nature of the process that underlies the socio-cognitive activity and 2) the nature of the target states tracked by that activity (cf. Zawidzki, forthcoming, ch. 1). Along the first axis, low-level indicates relatively automatic, involuntary and/or subconscious processes, whereas high-level amounts to relatively effortful, voluntary and/or conscious interpretation. Along the second axis, low-level refers to relatively simple, observable states and high-level to relatively complex, unobservable states. Arguably, relational mindreading is often relatively effortful, voluntary and accessible to consciousness. (Remember that our primary target is the *discursive* practice of interpreting, explicating, explaining actions). But there is nothing in the concept of relational mindreading *per se* that demands that it is a phenomenologically accessible, explicit form of reasoning about other people's mental states. On the other hand, relational mindreading occupies a specific end of the spectrum on the second axis.

The distinction between understanding someone's behavior as an expression of 'observable' mental states on the one hand, and interpreting her behavior being informed by 'unobservable' mental states on the other, need not be understood in terms of the inner-outer contrast, observable states reaching the bodily surface, unobservable states remaining 'hidden' in the head. Rather, the observable-unobservable dimension points to the degree in which

18 As for as I know, it was Goldman (2006) who first used this terminology in this context. Goldman describes 'low-level simulation' as "comparatively simple, automatic and below the level of consciousness" (p. 113) and contrasts it with 'high level simulation', which targets mental states of a relatively complex nature, such as propositional attitudes, is at least partly under voluntary control and/or has some degree of access to consciousness (p. 147).

a mental state type tracks a particular type of behavior. The prime example of an observable mental state is a motor-intention, a 'mental state' being specifically directed at the performance of a particular kind of action. Primitive emotions, such as disgust and fear, are also relatively observable states, matching specific kinds of facial expression and bodily posture. On the other side of the spectrum we find mental states with only tenuous connections to behavior types. Here the prime example is a state with propositional content. The relation between propositional states and their behavioral manifestations is one-to-many and many-to-one. Think of John again, having run out of milk. His reason for walking down the street is that he has run out of milk. But interpreting his action as being informed by the state that he has run out of milk only makes sense if we see it in the wider, temporally extended context of John's mind: his knowledge, intentions, plans, preferences, etc. Attributing to John solely the knowledge that he's run out of milk, devoid of any mental context (which arguably, is nonsensical in the first place) does not suggest any particular kind of behavior, and so every kind is a possibility. The same goes for understanding the goal of John's action. Sophisticated (longer term) goals are relatively unobservable compared to the immediate goals of simple motor intentions, such as grasping a cup or grabbing a doorknob.[19] Buying some milk in the supermarket is not something that manifests itself in any particular type of walking behavior. A mental state is unobservable to the extent in which its interpretation in relation to behavior requires a 'holistic' mental context (cf. Davidson 1970/2001a).[20]

Relational mindreading deals in relatively unobservable, relational states. Obviously these states cannot be regarded as 'inner' states: the agent (her head included) constitutes just one side of the relation. Relational mindreading allows the interpreter to relate another person to the world by propositional means. It allows for the attribution of genuine doxastic states and intentions, as well as the commitments and entitlements that these states carry on their shoulders. Relational mindreading thus permits us to participate in the game of giving and asking for reasons. It exhibits mastery of the inferential connections between our thinkings, sayings and doings, the kind of expertise that is required for criticizing someone's reasoning, holding him to his word or

19 Our focus is on such more sophisticated goals, not on the goals of motor intentions.

20 This holistic constraint on propositional attitude ascription has been recognized by many participants in the debate on high-level mindreading. See e.g. Heal (1996), Morton (1996, 2003), Currie and Sterelny (2000), Nichols and Stich (2003), Goldman (2006), Bermudez (2003; 2009), Zawidzki (2008). See chapter 5.3 for discussion.

asking him to explain his conduct. In its fullest form, relational mindreading amounts to explicit reasoning about someone's goals and reasons, which demands making inferences about unobservable, propositionally articulated mental states.

Thus, relational mindreading is a sophisticated form of human social understanding that enables us to perceive each other's relation to the world in non-representational, intentional, yet propositionally articulated ways. One of the reasons I have chosen the term 'mindreading' is that it is often used in the debate on folk psychology and social cognition as a theoretically neutral label for our capacity to attribute propositional attitudes (cf. Nichols and Stich 2003, p. 2; Apperly 2011, p. 3). Let me stress again that my primary target is an important *explanandum* in the debate: the capacity of normal adult human beings to interpret one another in terms of their goals and reasons. The technical notion of relational mindreading is meant to give a further characterization of this explanandum. My claim is that our capacity for propositional attitude ascription is relational rather than representational at base. It is not my primary aim to give a detailed *explanatory* account of the cognitive underpinnings of this capacity. The neutrality of the technical term 'mindreading' therefore serves my purposes quite well.

Some have criticized the use of the term, because the association with telepathy makes it appear as if our commonsense understanding of others is a rather mysterious, occult affair of 'gaining access' to their otherwise 'hidden' minds. As I use the term, however, it only serves to mark the contrast with other ways in which we are sensitive to the minds of others. Just as *seeing* the words on the page of a book is not the same as *reading* them, so perceiving other people's minds in their facial expressions, postures and bodily movements is not the same as interpreting them in inferentially articulated and truth-evaluable ways. There is nothing mysterious about reading the page of a book, and there need not be any hidden message to be uncovered between the lines. Likewise, reading other people's minds often takes the mundane form of simply listening to what they have to say about something, about their own actions, for example. In such cases, the propositional attitudes ascribed are enacted in the words interpreted – not encrypted, 'hidden' from view.

2.7 Conclusion

Human social practice hinges on the capacity of its participants to interpret each other in terms of goals and reasons for action. Most philosophers and psychologists have followed the BD-Model of action explanation and hold that commonsense goal-reason psychology is in fact belief-desire psychology. This chapter provided the essential ingredients for an alternative conception: the Relational Model of folk psychology. First, it introduced relational mindreading as a philosophical characterization of the psychology of default goal-reason attribution. On this account, representational mindreading, in terms of beliefs and desires, is an essentially secondary interpretation technique that enables us to manage and evaluate our discursive engagements with one another when social interaction becomes problematic. It also characterized the armchair tendency to parse ordinary goal-reason attribution in terms of beliefs and desires as a typical instance of the reflective fallacy, the fallacy of projecting one's sophisticated, reflective understanding of human social practice onto the psychology of unreflective participants of that practice. Finally, it gave a preview of how this armchair tendency might actually be explained by the present account, in terms of the specific social functions of belief-desire psychology mentioned above. The following chapters will be mixing these basic ingredients together, adding some further constituents as we go along. Toward the end, this will reveal the Relational Model of folk psychology as a genuine, and I hope plausible, alternative to the BD-Model.

The next chapter will focus exclusively on the *locus classicus* of functionalism in the philosophy of mind: Sellars's Myth of Jones. As we shall see, a functionalist treatment of folk psychology need not yield a representationalist picture of mind. It in fact invites a relational conception of folk psychological ascriptions.

Appendix: Motivating Reasons

The reasons that occupy a central place in commonsense goal-reason psychology are sometimes referred to as 'motivating' reasons. There are different philosophical accounts of what motivating reasons are, however, and of how they relate to other 'kinds' of reasons we give in folk psychological practice. This appendix provides a selective overview of these matters in relation to the present account.

There seem to be three ways in which we use the term 'reason' in everyday reason talk. First, we may be interested in the reasons *there are* for someone in certain circumstances to perform an action of a certain kind. Talk of reasons in such explicitly evaluative context serves the purpose of finding out whether acting thusly is the appropriate, prudent or right thing to do. Sometimes philosophers speak of 'normative' reasons in this context: reasons that make an action more or less favorable, right or wrong. Second, we may want to know the reasons *for which* someone acted. To see the difference with the first, normative use, notice that people may have good reasons for acting in a certain way, while not acting for those reasons, or contrary to those reasons. When we talk about reasons in this context, we are interested in the reasons that *motivated* the agent, even if the agent's reasons were bad reasons for acting in the way she did. Hence the term 'motivating' reasons. Third, there is a wider class of 'explanatory' reasons that may be mentioned in giving an explanation of why a certain event occurred or why a certain state of affairs obtains. Regarding the explanation of action, we can think of explanation in terms of character traits, habits, abilities, upbringing, situational factors, social roles, emotions, moods, etc. When explaining someone's action in this way, we may not be specifically interested in the reasons that motivated him. Rather, we seem to put his reasons, whatever they are, in a wider, personal context.

When discussing and criticizing the BD-Model of action explanation, it is crucial that we be clear about the kind of reason talk that we adopt. We might say that the BD-Model is concerned with intentional action *insofar as* it is explained in terms of motivating reasons. Thus the fact that we often cite an agent's traits, history, abilities or emotions in the course of explaining his action, *prima facie* does not pose a threat to the BD-Model of action explanation. For it need not claim that all explanations in terms of the wider class of explanatory reasons explain by invoking belief-desire pairs.[21] Within the

21 It seems that Smith goes beyond this characterization of the Humean account when he says that "once we see the central place occupied by Humean belief/desire explanations, we see

mindreading debate, this caveat has generally been ignored. Those inspired by the BD-Model have focused almost exclusively on belief-desire psychology at the expense of 'thicker' folk psychological explanations in terms of people's traits, motives, moods, habits, etc. (cf. Goldie 2007, see also Ratcliffe 2007).

That being said, it may be argued that explanation in terms of an agent's traits, history, abilities, etc. often does occur with implicit reference to the agent's motivating reasons. Consider Malle's (2001, 2004) account of action explanation, for example. Malle distinguishes two basic alternative strategies for the explanation of intentional action. Instead of citing an agent's reasons for action, interpreters may rather refer to 'the causal history of reasons', so as to describe background factors that triggered a reason for action (e.g. John cooked diner today because he was home early). The second alternative strategy makes use of 'enabling factors' in the explanation of intentional actions, situational factors that have a positive (enabling) or negative (disabling) influence on the performance of an action (e.g. Peter stayed up all night working because he had a lot of coffee). On Malle's account, explanation in terms of causal history (upbringing, character, social roles, moods, etc.) or enabling factors (abilities, emotions, physical fitness, etc.), *make sense* as explanations of intentional actions because they imply that the action was performed for a reason. If we know about someone's character, for example, we know, in general, *what kind* of reasons she is motivated by (see also Schueler, 2003). Thus, even if we are not certain about the specific reasons that motivated her, information about her character (or upbringing, etc.) gives us a pretty good idea where to look for them. Giving explanations in terms of enabling factors seems especially appropriate when it is the *manner in which*, rather than the reason for which, the action is performed that requires explanation (e.g. Peter stayed up working because he had a deadline; he stayed up working *all night* because he had a lot of coffee).

Explanations in terms of (moral) rules also fit this schema. Consider: he bought her a present because that's the way they celebrate birthdays around here. Or: she helped the old lady cross the street because that was the right thing to do. These explanations mention normative reasons for the types of action under consideration (it being her birthday today, an old lady trying to cross the street). But it may very well be the case that such explanations make sense as explanations of the particular action *tokens* in virtue of implicit refer-

that all the other explanations we give simply supplement this basic Humean story." (1998/2004, p. 156; emphasis added)

ence to the agent's motivating reasons. The explanations we provide in folk psychological practice in terms of traits, roles, rules, habits, abilities, emotions, etc. may be 'thicker' precisely to the extent that they reveal a *pattern* in the agent's motivating reasons. Rather than obviating reference to other people's motivating reasons, this wider class of explanatory reasons may actually *deepen* our understanding of their motivating reasons.

This brings us to the relation between motivating reasons and 'normative' reasons. The BD-Model of reason *explanation* tends to go accompanied by a 'Humean' account of motivating reasons *themselves*. Accordingly, motivating reasons consist of belief-desire pairs (cf. Smith 1987, 1994). Naturally, if motivating reasons *just are* BD-pairs, then attribution of such reasons must consist in the attribution of, *inter alia*, such pairs.

On this Humean account, motivating reasons and normative reasons are of a different ontological kind. Normative reasons are things that favor an action, e.g. worldly states of affairs or facts, such as it raining outside or the fact that holding up an umbrella will normally keep one dry (as reasons for carrying an umbrella). Motivating reasons, by contrast, are psychological states of the agent: appropriately structured belief-desire pairs such as desiring to stay dry and believing that carrying an umbrella is a good means to stay dry. It is because motivating reasons are considered to consist of BD-pairs that Humeans such as Smith hold that a motivating reason explanation explains *in virtue of* reference to a belief-desire pair that informs the action (see e.g. Smith 2003/2004)

A minority of philosophers of action has put forward a 'non-psychologistic' alternative for the Humean account (e.g. Dancy 2000, 2003; Bittner 2001; Schueler 2003; Stoutland 2007, Alvarez 2010). These philosophers claim that motivating reason explanations explain in virtue of citing that 'in the light of which' the agent acts. Suppose Jill runs out the door because the house is on fire. What it is, in the light of which she runs out the door, is the house being on fire, or the fact that the house is on fire.[22] The explanation of Jill's action succeeds by citing this fact – a fact about the worldly situation she responded to, *not* about her own psychological condition. Motivating reasons are thus

22 Proponents of the non-psychologistic alternative disagree about the ontology of reasons. Bittner (2001) and Stoutland (2007) think that reasons are concrete states of affairs or events, while Alvarez (2010) holds that reasons are abstract entities, viz. true propositions or facts. Dancy (2000) hovers between the two. On the one hand he thinks that reasons are concrete things like her distress or yesterday's bad weather (p. 115), and thus not abstract entities. But on the other hand he holds that people act for genuine reasons in error cases, committing him to the idea that reasons are things that can be the case (p. 145-151). But something that can be the case seems to be something abstract rather than concrete (see Alvarez 2010, p. 157).

conceived as the states of affairs that the agent's considerations *are about*, not as the psychological states or events that enable the agent to be intentionally directed towards these states of affairs.

On this alternative, motivating reasons are regarded as a subspecies of normative reasons. Accordingly, there is only one kind of reason for action (as opposed to several kinds of explanatory 'reasons why', such as traits, abilities, etc. – see above) that may be termed 'normative' or 'motivating' only relative to a context of interest. Thus, when deliberating what to do, an agent takes into account the 'normative' reasons she has for (not) performing a certain action. Making up her mind, she decides to perform the action for one or some of those reasons, and acts accordingly. Explaining her action then consists in picking out the 'motivating' reason she decided to act upon.

What is essential to this position is that reason explanations given from a second or third person point of view need to be such that they render the action rational from the point of view of the agent (see especially Dancy 2000, Schueler 2003 and Stoutland 2007).[23] This is exactly what the phrase 'reasons in the light of which' is supposed to capture. What explains Jill's running out the door is the fact that the house is on fire. It is this fact that makes Jill's running out the door a rational thing for her to do; it is a fact that could have figured in her deliberation leading to the decision that running out the door is the best thing for her to do. By contrast, citing her belief as an explanation for her action does not make sense from her point of view: the fact that she *believes* that the house is on fire, by itself, is no reason for running out the door. Compare this to the situation in which Jill checks the stove again because her fearful doubt that it might still be on keeps her from sleeping, or of a psychiatric patient who makes an appointment with his psychiatrist because he's hearing voices again. In these cases, the actions seem rational by the agents' lights in virtue of facts about their own psychological condition (the fact that

23 Non-psychologistic accounts of reason explanation in action theory show some parallels with simulation theory in the debate on folk psychology, especially Gordon's (1986) version of the simulation as performing an 'egocentric shift' (see chapter 4.2). Consider Stoutland's (2007) claim that "To be agent-centered, rational explanations must be formulated in terms of the agent's actions and the world as apprehended by her, but that is not at all the same as their being formulated in terms of her apprehension. What rational explanations require is not the agent's point of view but the world and her action as it is taken to be from that point of view." The Humean account, by contrast, suggests a Theory Theory account folk psychology. It does not deny we often give explanations in terms of normative reasons. But when we do, it says that these explanations are *themselves* to be explained with reference to, *inter alia,* the agent's beliefs and desires. (cf. Smith 2003/2004). The obvious candidate for such an explanation is a functionalist theory that specifies how beliefs, desires and other mental states are causally related to environment, behavior and each other.

her doubt keeps her from sleeping, or the fact that he is hearing voices again). But such cases are rather exceptional; most of the time people experience their actions as responses to things occurring around them, not in them.

On the Humean account, motivating reasons are belief-desire pairs and so motivating reason explanations explain in virtue of citing the belief-desire pair that constitutes the agent's motivating reason. On the non-psychologistic alternative, motivating reasons are a subspecies of normative reasons, namely those normative reasons in the light of which the agent acts, and so reason explanations explain by citing normative reasons. The question that needs to be asked at this point, however, is how we should understand the 'in light of which' clause from a second or third person point of view.

As argued in chapter 2.4, in order to regard the 'normative' reason for running out the door, i.e. that the house is on fire, as the reason that motivated Jill's action, it seems we need to see some kind of 'intentional connection' between Jill and her reason. We need to understand Jill as *being aware* of the fact that her house is on fire, and as getting into the state of *being motivated* to leave the house and run out the door. The clause 'in the light of which' is meant to capture the agent's intentional directedness towards the situation at hand; without it, a factive reason explanation could not explain her action. The question is whether this 'intentional connection' must be of representational kind, whether attributing a reason 'in the light of which' to another person requires the ascription of beliefs and desires. Philosophers of action tend to stay silent on matters of mindreading. But at times proponents of non-psychologistic accounts seem to accept the BD-Model of action explanation in the sense that belief-desire ascription forms a necessary *enabling factor* for reason explanation to succeed (Dancy 2000, p. 127; Alvarez 2010, p. 174).

There are obvious parallels between the non-psychologistic alternative to reason explanation and the present account. The relational mindreader relates the agent to her reason, which is typically not something psychological about the agent but rather something about the world outside the agent. On both accounts, 'motivating' reasons are things suited to figure in the deliberation of an agent or the 'co-deliberation' of the interpreter. Apart from this, it is not exactly clear how far the parallel goes. If in fact proponents of the non-psychologistic account agree with the BD-Model insofar as it holds that belief-desire ascription is necessary for reason explanations to explain from a second- or third-person view, then this of course would mark a clear difference. In the remainder of this appendix, I will focus on the accounts provided by Dancy (2000) and Alvarez (2010).

Non-psychologistic accounts contrast with the Humean account insofar

as they hold that it is in virtue of the agent's reason 'in the light of which' that a reason explanation explains, rather than in virtue of a Humean belief-desire pair of the agent. The 'in the light of which' clause (whether it be interpreted relationally or representationally) highlights a background condition for the explanation to go through; it is not part of the *explanans* itself.[24] The Relational Model could hold a middle position on this score, for it could say that the real *explanantia* of default reason explanation are relational mental states, states relating the agent to her reason. It is not the agent's reason that by itself constitutes the *explanans*, but rather that reason *in relation to the agent*. It could thus agree with the Humean account insofar as the latter holds that there is a psychological element in the *explanans*.

Alvarez (2010) thinks that we can only speak sensibly of motivating *reasons* in veridical cases, i.e., in which the agent has a true belief about that which motivates her. In error cases, she speaks of 'apparent reasons'. This, she argues, does justice to the fact that agents will normally retract their claim to having a reason when they find out that their action or intention was based on a false belief. This means that we can only give genuine reason explanations in veridical cases; in error cases we turn to Humean explanations that mention the agent's apparent reason in the content clause of her beliefs (p. 177-181). Alvarez thereby rejects Bernard Williams' principle that 'The difference between false and true beliefs on the agent's part cannot alter the form of the explanation which will be appropriate to his action.' (1981, p. 102)

The Relational Model is sympathetic towards Alvarez proposal, in that it stresses the important difference between interpretation in veridical cases (by means of relational mindreading) and interpretation in error cases (by means of representational mindreading). It can, however, retain an element of Williams' principle: that in both veridical and error cases there is a psychological element featuring in the *explanans*.

Dancy (2000) accepts Williams' principle, but rejects the Humean account of reason explanation. For Dancy, reason explanation from a second- or third-person perspective has an important and irreducibly first-person element.[25] He

24 Cf. Dancy (2000, p. 129) "That explanation specifies the features *in the light of which* the agent acted. It is required for this sort of explanation that those features be present to the agent's consciousness – indeed that they be somehow conceived as favouring the action; so there must always be a way of making room for this fact, in some relation to the explanation that runs from features as reasons to actions as response. It is not required, however, that the nature of the agent's consciousness itself either constitute, or even be part of, the *explanans*."

25 Dancy's account shows interesting parallels with Gordon's (e.g. 1986, 1995, 1996) 'radical' simulation theory. For Gordon, third-person interpretation is also essentially first-personal, in that it requires an 'egocentric shift' on the part of the interpreter, a 'transformative identification' with the agent to be interpreted. Dancy's 'appositional account' of belief ascription also has a lot in

says that 'The distinction between first and third person does not allow us to suppose that in the third-person case, there is a radical distinction between the psychologized and the non-psychologized forms of explanation, when there is no such radical difference in the first-person case.' (p. 135) Since the Humean account does not make sense from the first-person perspective, it needs to be abandoned from the second- or third-person perspective as well. Yet at the same time, because there is no radical distinction between veridical and error cases from the first-person perspective, we shouldn't allow for such a distinction from the second- or third-person perspective either. This commits Dancy to the claim that agents can act for genuine reasons even if their actions are based on false beliefs. Accordingly, reasons are things that may or may not be the case – things with the property of being suited to be the case – a metaphysical bullet that he is willing to bite (p. 145-151). It also moves him toward a non-causal account of reason explanation. For things that are only capable of being the case, he argues, cannot feature as *explanantia* in causal explanations (p. 161).

The first thing to notice is that the Relational Model can allow for a causal rendering of reason explanation: the attitude of the agent towards his reason – one of the *relata* of the relational mental state explaining the action – may be regarded as causally relevant in bringing about the action. And, of course, this causal rendering is also possible when the interpreter shifts to representational belief-desire psychology (see chapter 3 and 4).

Secondly, the strong symmetry Dancy claims exists between first-person acting on reasons on the one hand, and third-person reason explanation of action on the other, paradoxically suggests a full-blown representational rendering of third-person interpretation. For if the distinction between veridical and error cases is not allowed to enter into our account of third-person reason explanation, on the grounds of symmetry with the first-person perspective, then it seems we are committed to belief ascription or representational mindreading in both error *and veridical cases*.[26] This is paradoxical because such

common with Gordon's ascent routine approach: "…the appositional account. This hears 'He is doing it because he believes that p' as 'He is doing it because p, as he believes.' The 'as he believes' functions paratactically here, attaching itself to the 'p'. Again, it is not part of the specification of his reason, but is a comment on that reason, one that is required by the nature of the explanation that we are giving." (Dancy, 2000, p. 128-129) See chapter 4.2 for further discussion of Gordon's simulation theory.

26 It is not clear from the text of his (2000) where exactly Dancy stands on this matter. Some remarks, however, do point this direction. Consider: "There are, then, both factive and non-factive ways of laying out the considerations in the light of which the agent acted. If this is so, it seems to me that the difference between the factive and the non-factive cannot be of any real significance when it comes to the explanation of action." (p. 134)

representational understanding of the veridical case from the third-person perspective would in fact show an important *asymmetry* with the first-person perspective. For it seems that agents normally do not represent themselves to themselves when responding to a reason.

The Relational Model is not wedded to a strong symmetry thesis regarding the first-person and second- or third-person perspective. In fact, it suggests an important asymmetry. From the first-person perspective, an agent may just respond to a reason, without conceiving of that response as a relation between herself and her reason. But the second- or third-person interpreter who is attributing a reason to someone else must make room for the fact that it is *the agent* who is responding to that reason, and not she herself or someone else. The Relational Model accounts for this by having the interpreter conceive of the agent's response as a relation between that particular agent and her reason. It is this asymmetry that allows a psychological element to enter into relational mindreading, an element that is arguably absent in first-person experience of reasons.

3

Mindreading in Sellars's Myth of Jones

3.1 Introduction

Folk psychology often takes the form of what I have termed 'goal-reason psychology': of explicating the behavior of fellow human beings as being directed at goals in response to reasons. How should we further characterize this? The vast majority of philosophers and psychologists has adopted the BD-Model and holds that goal-reason psychology is in fact belief-desire psychology. The idea of commonsense understanding of other people's goals and reasons through the attribution of propositional, inferentially articulated, yet non-representational mental states has largely been ignored in the debate on social cognition.

As theorists reflecting on folk psychology, we are naturally inclined to adopt belief-desire terminology in characterizing the interpretation processes we try to understand. This inclination, so I suggested in chapter 2, stems from the fact that belief-desire psychology has precisely the function *of* reflecting on our goals and reasons when social understanding becomes problematic. The habit of parsing goals and reasons in terms of beliefs and desires upon reflection makes it very hard for philosophers and other theorists to get a relational stratum of goal-reason psychology into clear view. This chapter can be

regarded as an attempt to counter this reflective habit.

For this purpose, I will carefully reconsider Sellars's well-known 'Myth of Jones' in 'Empiricism and the Philosophy of Mind' (1956/1997). Sellars presents his myth primarily for expository reasons. It is a deliberately fictional story that aims at revealing certain important features of the conceptual structure of our folk psychology. Sellars is often credited with the dubitable honor of being the intellectual forefather of the Theory Theory of folk psychology. But what exactly did Jones teach the main characters in Sellars's myth, our 'Rylean ancestors'? The answer, I argue, is that he taught them how to engage in relational mindreading, *and nothing more*. Importantly, representational mindreading was still beyond their grasp when Jones's mindreading classes were over. I thus use Sellars's myth for my own expository ends; it is a very useful philosophical tool for revealing a relational conception of mindreading. At the same time, this puts the debate on folk psychology in a new light. What is widely regarded as the intellectual inspiration and conceptual basis for the representationalist theory theories of folk psychology actually provides us with nothing beyond a relational understanding of mind.

The next section presents an overview of Sellars's Myth of Jones, with a specific focus on the way in which Sellars lets Jones build his new 'theory of mind' on the foundations of certain pre-existent skills of our Rylean ancestors. In the third section, I will exploit Sellars's bootstrapping techniques as revealed in section 2. By carefully following through Sellars's strategy, it becomes vividly clear that Jonesian folk psychology cannot go beyond a relational understanding of each other's goals and reasons. The remainder of the chapter will put this conclusion in broader perspective. Section 4 homes in on the conceptual nature of the argument and distinguishes it from related empirical considerations. It also provides a friendly amendment to the working definition of relational mindreading presented in the previous chapter. Section 5 ends with a cautionary note on the alleged theoretical nature of mindreading on Sellars's account. In certain important respects, the Myth of Jones actually suggests a non-theoretical account folk psychological interpretation.

3.2 How Jones Taught Our Rylean Ancestors

Sellars's Myth of Jones in his seminal essay 'Empiricism and the Philosophy of Mind' (1956/1997; hereafter EPM) is often referred to as the original source of the Theory Theory of folk psychology. As explained in chapter 2.2, TT holds that our ordinary explanations, predictions, etc. of each other's thoughts and

actions rest on the (tacit) application of a commonsense theory that specifies how mental states are causally related to environmental conditions, observable behavior and other mental states. In certain crucial respects, however, the Myth of Jones does not lend credence to the details of TT. Most importantly, it does not support the dominant BD-Model of TT, according to which theorizing about other people's goals and reasons demands quantification over their beliefs and desires (see chapter 2). But this will have to wait until the next section. This section first gives a brief overview of the place of the Myth of Jones in EPM. It will then provide a more detailed discussion of the first part of Sellars's myth.

Sellars's ultimate target in EPM is the nature of certain forms of first-person, rather than second- or third-person ascription of mental states. He has in mind the judgments one can make about the contents of one's own experiences and thoughts. Sellars attacks traditional foundationalist accounts in epistemology with a special focus on so-called sense-datum theories, still very popular at the time he wrote the essay. According to foundationalist theories, our judgments about our own experiences and thoughts have the special epistemic status of being a warrant of absolute certainty for our further epistemic concerns. On this typically Cartesian picture, our epistemic access to our own thoughts and experiences is completely transparent; thoughts and experiences are self-authenticating episodes such that the mere *having* of an experience or thought entails *knowing that* one has the experience or thought. Accordingly, the mere act of experiencing or thinking provides one with incorrigible first-person knowledge, an epistemic *datum* or *Given*, as Sellars calls it, on which all other knowledge is founded. For a number of reasons, not to be discussed here, Sellars thinks that the idea of such an epistemic Given is deeply problematic, a philosophical myth. The Myth of Jones is Sellars's antidote against the myth of the Given. It is "a myth to kill a myth" (EPM, §63), a piece of anthropological fiction that purports to sketch an alternative picture of the first-person epistemology of thought and experience. Crucially, it starts by telling a story about second- and third-person ascription. It is this part of the myth that will occupy us further on.

The Myth of Jones begins with a description of a prehistoric community of humans Sellars calls our 'Rylean ancestors'. They are 'Rylean' in the sense that they do not have any conception of inner mental life, of their own in particular.[27] Their concept of mind is confined to public displays of intention-

27 Obviously, this Rylean conception of mind is modeled on Gilbert Ryle's (1949) behaviorist treatment of the concept of mind.

ality, such as the making of overt linguistic utterances. One day, a genius called Jones appears who comes up with a 'theory of mind', according to which overt utterances are caused by inner 'thoughts'. The beauty of the theory is its simplicity. For these theoretical posits Jones calls 'thoughts' are modeled on overt utterances themselves, an understanding of which the Ryleans have already mastered. Building on this understanding, Jones is able to teach the Ryleans his new theory by analogy. He thus tells them that when they speak, this is the result of something analogous going on inside them, a process of 'inner speech'. He also tells them that such inner episodes can occur without overt linguistic manifestation. With the core of the theory in place, Jones trains his Rylean students to make correct ascriptions of thoughts to others; he shows them how ascription of the thought that *p* to an agent is correct when, roughly, the overt utterance that *p* would make proper sense for the agent in that context. He then teaches them to apply this newly acquired skill to themselves. First they make self-ascriptions inferentially, on the basis of indirect behavioral evidence, behavior caused by the thought ascribed (e.g. their own overt utterances). After a lot of practice, however, the Ryleans acquire the disposition to self-ascribe as a direct response to simply *having* the thought, thus 'bypassing' the inference from the behavior that it causes. As a result of their training the Ryleans have learned to respond to the thought that *p* by *reporting* rather than inferring that they are themselves thinking that *p*. As it turns out, this non-inferential form of self-ascription is normally a very reliable indicator of the presence of the thoughts ascribed, more reliable, in fact, than most ascriptions by *others*. At the end of the story, therefore, "our ancestors begin to speak of the privileged access each of us has to his own thoughts" (§59) – privileged, but by no means infallible access, as traditional foundationalist accounts have it. In the last section of the essay Sellars proposes a similar treatment for first-person reports on sense-impressions, thus providing an alternative for the 'sense-data' of sense-datum theories. For our purposes, however, Sellars's treatment of thoughts is of particular interest, since it specifically concerns attribution of mental states with propositional content.

Sellars main concern, then, is to explain the privileged status of first-person judgments about the contents of one's own thoughts without falling victim to the myth of the Given. Using the notion of a theory as a model for his account of such folk psychological claims, he hopes to have shown how, like scientific theories, these claims figure *in* a rational, self-correcting social practice, rather than as its foundation. Thus, the Myth of Jones

> helps us understand that concepts pertaining to such inner epi-

> sodes as thoughts are primarily and essentially *intersubjective* […] and that the reporting role of these concepts – the fact that each of us has a privileged access to his thoughts – constitutes a dimension of the use of these concepts which is *built on* and *presupposes* this intersubjective status. (§59, emphasis in original)

This last quote reveals that one of Sellars's primary aims is to show that our concepts of 'inner' episodes such as that of thinking that p can be derived from our concepts of publically observable occurrences such as that of saying that p.[28] In later work, he would refer to the conceptual material out of which to reconstruct our commonsense folk psychological notions as 'Verbal Behaviorism' (VB). In 'Meaning as Functional Classification' (1974/2007; hereafter MFC) he explains: "According to VB, thinking 'that-*p*', where this means 'having the thought occur to one that-*p*,' has as its *primary* sense *saying* 'p'; and a *secondary* sense in which it stands for a short term proximate propensity to say 'p'." (p. 83, emphasis in original)[29] Importantly, the VB model does not allow for the idea that thinking is a stream of consecutive inner episodes that may, but need not be, *expressed* in overt speech. According to the VB model, all thinking is *thinking-out-loud*, as *we* would put it. The VB conception of propositional mental content is confined to having the propensity or disposition to say certain things under certain circumstances; it conceives of the process of thinking as rapid and complex shifts of such dispositions.

The pre-Jonesian Ryleans are such verbal behaviorists; their appreciation of each other's mental lives is confined to an understanding of each other's propensities to display certain kinds of behavior, linguistic acts in particular. This by itself is already a very sophisticated skill they possess. It requires that they have mastered a language, that they know when it is appropriate to say certain things and how to act on certain sayings, what follows from what one has said, what it follows from and what it is (in)compatible with, etc. The Ryleans already have a highly developed conception of intentionality, in the form of the aboutness of their sayings and the directedness of their doings.

This is an important point. Unlike some forms of logical behaviorism, Sellars's Verbal Behaviorism does not aim at *reducing* intentionality to behavioral dispositions.[30] Rather, the verbal behaviorist conception of intentionality

28 For a discussion of the sense in which thoughts are conceived as 'inner' episodes on Sellars's story, see section 4.

29 A considerably revised version of this paper appeared in Naturalism and Ontology (1980, ch. 4). I will be referring to the original version, reprinted in Scharp and Brandom (2007).

30 By calling the protagonists of his story 'our Ryleans ancestors', Sellars's makes it clear that

the Ryleans already possess is *confined to* behavior, behavior that is already conceived as intrinsically intentional, as instances of *acting* and *saying*.[31] Another point worth stressing is that Sellars is making a *conceptual* claim here, not an ontological one. He is not saying that the pre-Jonesian Ryleans *did not have inner mental lives*. They already instantiated inner episodes of thinking, episodes that, on Sellars's story, caused their overt utterances and intentional actions. Verbal Behaviorism is an account of the Ryleans' *conception* of the mental; the point of the Myth of Jones is to show how the Ryleans can be bootstrapped into a more sophisticated way of *thinking about* thinking, a way that is actually explanatorily superior to Verbal Behaviorism, in approximating the true causes of intentional behavior. Sellars's strategy is to treat the VB conception of thought as thinking-out-loud is *conceptually* prior to our conception of inner thought episodes and to analyze the latter by analogy with the former. But this is compatible with the idea that inner thoughts are *causally* prior to the occurrence of meaningful speech. That, in any case, is what Jones teaches the Ryleans, and what Sellars regards as a first step toward an accurate understanding of mind.

Following Sellars's methodological behaviorist strategy, pre-Jonesian folk psychology is thus to be understood as rather sophisticated in being sensitive to full-fledged propositional forms of intentionality, yet rather primitive in being sensitive only to public displays of such forms of intentionality. The challenge for Sellars is to show how this Rylean understanding of thinking as thinking-out-loud can be lifted up to a genuine appreciation of thinking as silent inner episodes causing overt speech, and to do so, of course, without presupposing any reference to such inner episodes in pre-Jonesian folk psychology. In order to meet this challenge, Sellars starts by adding two ingredients to the linguistic repertoire of the Ryleans, skills that by themselves do not rest on an understanding of thoughts as inner episodes, but which, when suitably combined, can yield such an understanding. These two ingredients are the capacities to engage in *semantical* and *theoretical* discourse.[32]

---

he doesn't interpret Ryle (1949) as having been aiming for a reductive behaviorist analysis of mind either. I think this is a fair interpretation of Ryle. Cf. Schwitzgebel (2002).

31 Cf. 'Language as Thought and Communication' (1969/2007, p. 80): "Thus, at the primary level, instead of analyzing the intentionality or aboutness of verbal behavior in terms of its expressing or being used to express classically conceived thoughts or beliefs, we should recognize that this verbal behavior is *already thinking in its own right,* and its intentionality or aboutness is simply the appropriateness of classifying it in terms which relate to the linguistic behavior of the group to which one belongs." (emphasis in original)

32 Talk of 'addition' of semantical and theoretical discourse here should be understood primarily in the *methodological* sense of not being presupposed by the characterization of the 'original' Rylean language. It need not be interpreted as a phylo- or ontogenetic hypothesis concerning

First, the Rylean language "would have to be enriched with the fundamental resources of semantical discourse." (EPM, §49) The Ryleans, that is to say, need to have a meta-language; they need to be able to speak *about* their first-order linguistic utterances. The meta-language Sellars adds to the Rylean language gives expression to a form of functional role semantics. Building on their capacity to *use* their language, Sellars grants the Ryleans the capacity to talk *about* their language use, i.e. to state the *function* of a certain expression in their linguistic practice according to the way they use that expression. Sellars characterizes the function of linguistic acts in terms of three kinds of norm-governed behavioral uniformities: (i) non-inferential language entry transitions from worldly features in perceptual situations to perceptual claims, (ii) inferential intra-linguistic transitions and (iii) non-inferential language departure transitions from avowals of intention to the corresponding actions (see appendix). Mastery of a language requires that one displays the behavioral patterns specified by the entry/inference/exit rules of the expressions belonging to that language.[33]

In order to say in descriptive terms what the function of a given expression in a language is, one would have to explicitly state all the entry/inference/exit rules that specify its meaning – crudely, in what perceptual situations (not) to use the expression, what (not) to infer from it or to infer it from and what (not) to do upon drawing practical conclusions related to the expression. Fortunately, this is not something the Ryleans need to be able to do in order

human linguistic practice (see section 3.4 for discussion). Thus in MFC (p. 89), Sellars comments: "It would be a mistake to suppose that a language is learned as a layer cake is constructed: first the object language, then a meta-language, then a meta-meta-language, etc. […] The language learner gropes in all these dimensions simultaneously. And each level of achievement is more accurately pictured as a falling of things belonging to different dimensions into place, rather than an addition of a new story to a building."

33 According to Sellars, being a full-fledged member of a linguistic community not only requires that one behaves in conformity with the rules of the language, but also that one can *obey* the rules, i.e. that one is able to think of oneself *as* complying to the rules, i.e. that one has an understanding of these rules as such. A full treatment of Sellars's account of linguistic behavior goes beyond the scope of this chapter. In essential outline, the idea is that members of a linguistic community train their children and correct each other so as to display the behavioral patterns that there *ought to be* in conformity with the rules of the language, which requires that they themselves as teacher or critic understand that one *ought to do* as the ought-to-be rules dictate. In teaching their children or criticizing each other, language users must therefore have a grasp of the rules themselves. Importantly, the child learning a language, or the adult speaking 'candidly and spontaneously', is not *acting intentionally* in accordance with the rules, she merely displays the pattern-governed behavior she has been trained to exhibit (see section 3.3). Sellars accounts for the normativity of the behavioral uniformities exhibited in linguistic practice by placing them in a deontic social context in which trainers, by obeying the ought-to-do's, condition the trainees into behaving in conformity with the ought-to-be's. See e.g. 'Some Reflections on Language Games' (1954/2007) and 'Language as Thought and as Communication' (1969/2007). For a comprehensive introduction to Sellars's account of rule-governed behavior, see Rosenberg (2004/2007), deVries (2005, ch. 2) and O'Shea (2007, ch. 4).

to grasp the idea that their utterances have a function in linguistic practice. They already speak the language; their newly acquired meta-language can ride piggyback on their first-order language use. What Sellars needs to add to their linguistic competence is, in effect, only a relatively simple 'procedure' that enables them to comment *upon* this competence (see appendix for discussion). Roughly, this procedure tells them that the function of an expression E in their language is exhibited by their proper use of E in actual linguistic practice. This function can be made explicit on demand and to the relevant extent, by following through and writing down the entry/inference/exit rules of E exhibited in proper use. Yet the fact that E has a function that is describable in this way and exhibited in its proper use is something that the Ryleans are now able to make intelligible to themselves. This first addition to the Rylean language, then, endows the Ryleans with a rudimentary functionalist conception of their own linguistic utterances, i.e. a conception according to which their linguistic utterances have a functional role that can be characterized in terms of entry/inference/exit transitions. Importantly, this conception does not appear to presuppose the concept of thought as inner episode. If an understanding of (being disposed to) thinking-out-loud does not presuppose this concept, then neither does *talking about* such understanding.

The second addition concerns the enrichment of the Rylean language with *theoretical* discourse:

> Thus we may suppose these language-using animals to elaborate, without methodological sophistication, crude, sketchy, and vague theories to explain why things which are similar in their observable properties differ in their causal properties, and things which are similar in their causal properties differ in their observable properties. (EPM, §52)

We should endow the Ryleans, that is, with the capacity to explain and predict the dispositions of observable entities by positing unobservable, i.e. *theoretical* entities: underlying non-dispositional states with certain causal properties. Thus, to use two of David Armstrong's (1980) examples, we should allow the Ryleans to theorize about the internal structure of a vase that causes it to break when pushed off the table, or about the microscopic properties that explain the poisonous effect of some substance upon ingestion. Theorizing in this sense, Sellars proposes, consists in finding an appropriate *model* for the theoretical entities posited, "i.e. to describe a domain of familiar objects behaving in familiar ways such that we can see how the phenomena to be explained would arise if they consisted of this sort of thing." (EPM, §51) The

Ryleans could for example use some of the macroscopic properties of, say, termites – e.g. their capacity to eat their way through wood – as a model for the causal roles of the theoretical entities they posit to explain the toxicity of the poisonous substance – e.g. poison particles 'eating' their way through the intestinal walls causing internal bleeding and, eventually, death. Importantly, a model is always accompanied "by a commentary which *qualifies* or *limits* – but not precisely nor in all respects – the analogy between the familiar objects and the entities which are being introduced by the theory." (*ibid.*, emphasis in original) To carry on with our example, the poison particles could be considered like termites insofar as they make holes in the intestinal walls, but not to the extent that they, say, have jaws and walk on six legs. Again, it should be realized that this capacity to theorize *simpliciter* also does not presuppose an understanding of thoughts as inner episodes. After all, theorizing about such things as poisonous substances or vases does not seem to touch on matters of intentionality at all, let alone in covert form.

With the addition of semantical and theoretical discourse to their linguistic practice, the Ryleans are finally in a position to appreciate the teachings of genius Jones. For Jones can now simply teach them how to combine their meta-linguistic and theoretical skills into a *new* capacity to theorize about their minds. First, he shows them how to apply their capacity for theoretical thinking to the objects of their own meta-linguistic discourse. He tells them that their disposition to use an expression to the effect that *p* in certain conditions can be explained by positing an unobservable inner episode of 'thought' that causes the overt utterance under those circumstances. Jones, in other words, "develops a *theory* according to which overt utterances are but the culmination of a process which begins with certain inner episodes." (EPM, §56)

Secondly, he exploits their meta-linguistic skills by having them characterize the causal roles of thoughts in terms of the functional roles of their linguistic utterances, thus teaching them, in effect, how to model thoughts on the overt utterances they serve to explain – i.e. as thinking *that p*. Differently put, Jones's model for these episodes "*is that of overt verbal behavior itself. In other words, using the language of the model, the theory is to the effect that overt verbal behavior is the culmination of a process which begins with 'inner speech'.*" (§56, emphasis in original) Jones's theory uses overt speech as a model; it also includes a commentary on the model that qualifies the analogy being proposed. Inner thoughts are like speech acts in being *semantically evaluable*, i.e. in *meaning* or *being about* something, but not, the commentary hastens to add, in involving "the wagging of a hidden tongue" or the production of any sounds (§57).

Modeling inner thoughts on overt linguistic utterances, the Ryleans are

able to infer the presence of a thought that *p* in someone whenever an utterance of this person to the effect that *p* would make proper sense of her behavior under the circumstances. And this, in turn, helps the Ryleans to appreciate "the fact that [their] fellow men behave intelligibly not only when their conduct is threaded on a string of overt verbal episodes – that is to say, as *we* would put it, when they 'think out loud' – but also when no detectable verbal output is present." (§56, emphasis in original) For now that they understand that someone's disposition to think-out-loud that *p* when she for example observes a certain state of affairs is causally mediated by the state of thinking-to-herself that *p*, it is but a short step to coming to realize that her observing might cause her thinking without the latter giving rise to overt speech. Similarly, seeing someone's disposition to e.g. act to the effect that *p* when she intends-out-loud to p as being instigated by an intending-to-herself to *p*, the Ryleans can come to appreciate the fact that such silent intendings might directly cause the corresponding actions, without intervening utterances.

As a result of Jones's teachings, the Ryleans have evolved from verbal behaviorists to 'verbal' functionalists. Applying the functional classification of their linguistic utterances to the theoretical entities posited to explain each other's utterances, the Ryleans have bootstrapped themselves into a rudimentary functionalist understanding of each other's displays of intentionality. The concept and contentfulness of inner thought is modeled on the concept and meaningfulness of public speech. At the same time, however, the occurrence of public speech is explained by the occurrence of inner thought. On Sellars's account, this puts the conceptual and causal dependencies exactly right.

This concludes our survey of Sellars's Myth of Jones. At this point in the story, the Ryleans have learned how to attribute thoughts to one another in the service of making sense of their behavior. As we have seen in the overview at the beginning of this section, in the final stage Jones conditions the Ryleans to apply their new theory of mind non-inferentially to themselves, thus giving them very reliable and even privileged access to their own thoughts (and, in the last paragraphs of EPM, to their own experiences). This final stage is the crux of the story within the context of the overall aim of EPM, i.e. to kill the myth of the Given. The present focus lies elsewhere, however. Our discussion of the first part of Sellars's myth has revealed how Jones taught the Ryleans a theory of mind by having them model their conception of inner thoughts on their verbal behaviorist conception of public speech. The question before us is what notion of mind Jones thereby endows the Ryleans with, in specific: whether it is a genuinely representational notion of mind.

3.3 What Jones Taught Our Rylean Ancestors

After Jones's teaching sessions are over, the Ryleans are genuine mindreaders, in the sense specified in chapter 2: they understand each other's behavior in terms of unobservable, propositionally articulated thoughts. The term 'thought' that Jones introduces into his new theory is actually a generic notion that covers different forms of 'saying-to-oneself' that the Ryleans already 'thought-out-loud' before Jones's arrival. Consider the distinction between claims stating or inferring how it is with the world and practical avowals specifying how the world is to be changed as a result of one's actions. Interpretation of each other's utterances in pre-Jonesian social practice proceeded in conformity with the entry/inference/exit rules that specify the appropriate use in their language. The pre-Jonesian Ryleans were quite good in assessing the non-inferential responsiveness their compatriots exhibited in making perceptual claims. Based on prior experience of others' dispositions to respond (in) appropriately to the perceptual scene, for example, the Ryleans displayed a remarkable sensitivity to the (un)reliability of each other's overt perceptual judgments. They knew whom they could treat as a reliable indicator of which aspect of the world. Having mastered the inferential roles of the expressions belonging to their language, they also had a keen eye for the (im)propriety of the inferences people overtly made. On the language exit side, they displayed a nuanced appreciation of each other's 'intendings-out-loud', knowing which such practical avowals were (in)compatible with which others, for example. And keeping track of each other's (in)appropriate non-inferential responsiveness to such overt intendings by performing the corresponding actions, they knew whose practical avowals to perform what kind of actions they could rely on.

But interpretation in the pre-Jonesian community was not restricted to the making of simple factual claims and practical avowals. Sellars carefully notices that the Ryleans were already capable of making and understanding subtle use of the subjunctive conditional.[34] They had the resources to wield dispositional concepts, which was manifested in their ability to talk about what would happen and what people would do in hypothetical situations. It

34 See EPM, §48: "Imagine a stage in pre-history in which humans are limited to what I shall call a Rylean language, a language of which the fundamental descriptive vocabulary speaks of public properties of public objects located in Space and enduring through Time. Let me hasten to add that it is also Rylean in that although its basic resources are limited [...] its total expressive power is very great. For it makes subtle use not only of the elementary logical operations of conjunction, disjunction, negation, and quantification, but especially of the subjunctive conditional."

is this capacity for counterfactual thinking that allowed them to start using their functionalist meta-language (characterizing their first-order language use in terms of what would be appropriate to say or do under certain circumstances) and to engage in theorizing (explaining dispositional properties in terms of underlying non-dispositional states), already before the arrival of Jones. Building on their capacity for counterfactual thinking, it is not unlikely that they also already had come to terms with rudimentary notions of possibility and necessity. This would have manifested itself in their intelligent use and interpretation of terms like 'perhaps', 'certainly', 'possibly, 'probably', etc. Accordingly, they would know how to differentiate between the significance of someone's claiming that *p* is perhaps the case and her stating that *p* is certainly the case, or someone's predicting that *p* will occur and her assessment that it might happen.

There is also no reason to think that pre-Jonesian appreciation of each other's sayings was restricted to the purely verbal aspects of each other's overt sayings. Plausibly, the Ryleans were sensitive to certain emotional cues present in one another's linguistic performances, such as tone of voice, facial expression and bodily posture. They probably knew how to differentiate between the epistemic merits of a doubtful answer and a confident response. Nor should we assume that their interpretation of each other only concerned the more 'cognitive' varieties of 'thinking-out-loud', such as claiming, affirming, denying, supposing that or wondering whether *p*. More emotional counterparts, such as regretting, fearing, hoping or being angry about the fact that p would also have been within their interpretative grasp. They need not have experienced any problems in responding appropriately and differentially to someone's dreadful whisper, as opposed to, for example, her enthusiastic announcement that *p*.

In sum, the pre-Jonesian Ryleans already treated each other as more or less reliable indicators – each other's sayings as more or less reliable indications – of (emotionally significant) situations or events reported on or to be effectuated by their own actions. Their linguistic sophistication allowed them to understand some of these sayings as concerning counterfactual scenarios and probably also as stating the (im)possibility or (im)probability of (past, present or future) states of affairs or events talked about.

Jones could exploit all these pre-existent interpretation skills when he showed the Ryleans how to model the theoretical entities of his revolutionary theory on each other's overt sayings. The question we should ask at this point, however, is whether, in their *post*-Jonesian stage, the Ryleans are sensitive to the *representational* dimension of all these different forms of thinking-to-oneself

they have now come to appreciate. Do their Jonesian mindreading skills allow them to appreciate the fact that the contents of a person's thoughts specify the world *as it appears to this individual person*, as opposed to the way the world presents itself to them? Do they understand that other people's thoughts can be true or false, in the sense that their thoughts articulate a worldview that (mis)represents the world? Our discussion of Jones's bootstrapping techniques in the previous section reveals in a fairly straightforward fashion that this cannot be the case.

In their pre-Jonesian stage, the Ryleans had no inkling of the private mental lives of others. They were verbal behaviorists: their understanding of each other's minds was confined to public manifestations of intentionality, such as overt sayings. As we have seen in the previous section, the addition, before the arrival of Jones, of semantical and theoretical discourse to their linguistic repertoire, did *not* change the Ryleans' verbal behaviorist conception of each other. When Jones finally enters the scene, *he doesn't add anything to their existing skills*. He simply, ingeniously, combines their skills into a new way of reading the minds of their fellow men and women. He shows them how to apply their theoretical skills to their own linguistic performances by using the functionalist meta-language they had already mastered to characterize the theoretical entities posited by this new theory. The functional specifications of these inner episodes are thus modeled on the functional specifications of their *pre-Jonesian* language, specifications which are given entirely in terms of the state of the world *as it is publically conceived*. Jones's theory enables the Ryleans to interpret overt action as being caused by inner intentional episodes, but the contents of these episodes are still characterized *entirely* with reference to the public world. In their pre-Jonesian state of mind, the Ryleans interpreted each other's utterances solely with reference to the world – things that (might have) happened in the past, states of affairs that (possibly) obtain at present, events that will (probably) occur in the future, etc. The whole idea of there being *different views* on these worldly states of affairs and events (or their likelihood, etc.) eluded them. The only thing that Jones does upon his arrival is, in effect, to teach the Ryleans how to exploit their understanding of each other's overt speech in the service of applying his new theory of thought. By doing so, Jones does not miraculously endow them with a representational understanding of intentionality; Jonesian mindreading is still entirely bound to the public sphere.

Let us go over this more slowly. There is an ambiguity in the notion of mental representation that makes the present point particularly difficult to understand. In one sense of the term, the Ryleans already had a concept of

mental representation before the arrival of Jones, a concept Jones teaches them to apply to the silent counterparts of their overt utterances.

Consider once more the pre-Jonesian stage in Sellars's myth. Already at this point in the story, the Ryleans are able to talk about proper use of their language in semantical discourse. They understand that proper use is a matter of how an utterance of a particular type is supposed to function within the game of giving and asking for reasons: in response to which perceptual circumstances one is supposed or (not) allowed to make certain perceptual claims (entry transitions), which premises should or may (not) elicit which conclusions (inference transitions), and which actions one ought or may (not) perform in response to which practical conclusions (exit transitions). By commenting on proper entry transitions, they can say what a certain perceptual statement *ought to track* in the immediate environment; by criticizing each other's inferences, they can reveal what their (perceptual) statements *ought to say* about the wider, non-perceivable world; by monitoring each other's practical avowals to act, they can explicate what such avowals *ought to tell* about the way the world is to be changed as a result of the intended action.

Engaging in semantical discourse in this manner, the Ryleans are in effect using a respectable notion of representation. Appropriate use of an utterance is specified by what, according to its entry/inference/exit rules, *it is supposed to indicate about the world*, both the perceptual environment (entry transitions) as well as the world at large – past, present and future (inference and exit transitions). The sounds their compatriots make are treated as linguistic *vehicles* whose *function* it is to map onto certain aspects of the world, and this is something the Ryleans are able to say in so many words in their semantical meta-language.

Let us term this the 'functional role' or FR-notion of mental representation.[35] Some readers may be reminded of so-called 'teleosemantic' theories of mental representation. In general, teleosemantic theories hold that something is a representation in virtue of the function it performs in a representational system. On Dretske's (1988) account, for example, a state (lawfully) co-varying with some environmental condition becomes a representation of that condition when it acquires the function of co-varying with it in causing certain behavior. And on Millikan's proposal (e.g. 1984), the function of the representational vehicle is characterized in terms of its role in guiding co-operating

35 Thus, overt utterances (and, by analogy, their inner counterparts) are considered representational vehicles in virtue of fact that they occupy a functional role specifiable in terms of the entry/inference/exit rules that capture their proper use in linguistic practice.

consumer devices in the performance of their proper functions by mapping reliably enough on certain environmental features. For our purposes, the details of and differences between these (and other) teleosemantic accounts are not important. The point of bringing them to mind is to direct attention to the fact that pre-Jonesian interpretation allows for a rudimentary, folk notion of representation that is in important ways analogous to the teleosemantic notion.[36]

What unites all teleosemantic theories, as Millikan (2004, ch. 5) points out, is their story about what it is for a representation to *mis*represent. Misrepresentation is explained in terms of a representational vehicle's *mal*-functioning in some way or other. False representations fail to perform their proper function and thereby fail to represent. The failure of false representations to represent should *not* be understood in terms of a *mismatch* of some sort between *what is represented by the false representation* on the one hand and the way the world actually is on the other. On teleosemantic accounts, Millikan concludes, "'What is represented' by a false representation is indeed 'something that does not exist,' *because a false representation represents nothing at all*." (p. 65, emphasis added)

And so it is on the Ryleans' pre-Jonesian understanding of 'overt' mental representations. At this stage in their development, incorrect use of their language can only be regarded as a failed attempt to perform a linguistic act. If a speaker amongst them claims that *not-p* when it is in fact the case that *p*, she is interpreted as merely *having failed to assert that p*, not *also* as having successfully expressed an inner world-directed thought that *not-p*. Errors of this kind are taken as evidence of a misalignment with reality, a glitch in linguistic performance, a malfunctioning of the speaker's linguistic skills. She *should* have said that *p* under these circumstances; her failure to comply with the norms of their linguistic practice may be responded to by giving her incredulous stares, by firmly making claims to the contrary, or perhaps by beating her with sticks. Whatever sanctions her fellow Ryleans subject her to, their intention cannot be to *change her views about* whether or not *p*. What they intend to change, if anything at all, is her misuse of their language, period. As far as the pre-Jonesian Ryleans are concerned, there is nothing behind her error that could

36 There is also a sharp disanalogy, however. Teleosemantic theories are designed to give a naturalistic *explanation* of mental content. By contrast, the pre-existent Rylean meta-linguistic skills that Jones exploits for teaching them his theory merely serve the practical purpose of *classifying* each other's utterances in order to criticize, correct or praise their linguistic performances (see appendix for further discussion). Yet the notion of representation involved, whether in providing reductive explanations or in making mere classifications, is basically the same.

make her utterance rational *from her individual point of view*.

The Ryleans take each other's thoughts to indicate what they ought to tell them about the world. If the utterance is uttered in the appropriate circumstances, the interpreter takes it to be about what it is supposed to indicate; if it is used inappropriately, the interpreter takes it to have failed to indicate. Misalignment with the world is not also interpreted as a sign of successful expression of 'something' else – the speaker's private take on the world. On the pre-Jonesian account, a linguistic utterance that misrepresents merely fails to indicate what it ought to. That's all there is to it. Indication could be taken quite literally as 'pointing' towards some feature or features in the world.[37] For the pre-Jonesian Ryleans, talking and thinking is just a sophisticated, propositionally articulated way of pointing. The pointing vehicle (the utterance or thought) can be considered a representation insofar as it has the function of aligning with certain features in the world, a function that can be described in terms of the entry/inference/exit rules, rules which the Ryleans can articulate to a certain extent in their semantic meta-language. Failure of a pointing gesture to perform its function is regarded as just that: a failure. It is like a road sign pointing in the wrong direction: it simply misdirects the interpreter.

On a mere FR-understanding of representation, then, misrepresentation is mere malfunction leading to misalignment. Pre-Jonesian mastery of the FR-notion of mental representation by itself does not yield representational mindreading, in the sense I have been using the term: it does not amount to an understanding of the 'overt' mental representations attributed to others *as giving expression to their own, subjective and possibly mistaken views on the world*. As we have seen, Jones does not, upon his arrival, add anything to the Ryleans' pre-existent ability to classify their language use in terms of FR-representations. He merely exploits their meta-linguistic skills for specifying the contents of the inner thoughts he introduces to explain their behavior.

In their post-Jonesian stage, the Ryleans are just as baffled about their compatriot's claim that not-p when in fact *p*: *she simply failed to claim-to-herself that p*. What they still cannot understand is that she thereby *succeeded in rep-*

37 My use of the term 'indication' here differs from Dretske's (1988). Indication on Dretske's account is mere law-like dependency between indicator (I) and feature indicated (F), such that F occurs whenever I occurs. The dependency either exists or it doesn't; when it does there is an indication relation, when it doesn't there is not. Hence, on Dretske's use of the term, an indicator cannot sensibly be said to *fail* to indicate. On a Jonesian reading, however, the indicator is a purposeful, rule-based (covert) linguistic act. Here, failure to indicate as one is supposed to makes perfect sense. Yet, as on Dretske's account, indication is a relational notion. If a speaker uses an utterance according to its function, she indicates something about the world; if she uses an utterance inappropriately, she fails to indicate. Hence, *if* she indicates, the indication implies the (past, present or future) existence of the indicated.

resenting the world to herself as if it were the case that *not-p*. The inner episodes the educated Ryleans attribute to one another are treated as *silent* events all right, but not as *private* ones. Their mental ascriptions still cannot make room for the fact that the contents attributed may give expression to the way one individual person conceives of the world, as opposed to the way the world is – publically, so to speak. The *privileged* access that the Ryleans grant each other with respect to their own thoughts only reflects the fact that first-person reports tend to be more reliable indicators of the presence of the thoughts ascribed than second-or third person ascriptions; it does not manifest appreciation of the fact that what one has privileged access to is how the world appears specifically to oneself and not necessarily to anyone else.

Representational mindreading is not an option for the post-Jonesian Ryleans. When confronted with a false claim or an unrealistic intention of their fellow men and women, they can only regard it as a failed attempt to say or bring something it about the public world, not *also* as a successful expression of their own private take on the world. In this respect, they are just like their former pre-Jonesian selves. To their ears, a false statement is a sign of mere misalignment. To their eyes, a misinformed or misguided action is a matter of mere malfunctioning, like a missile missing its target. In such cases, their interpretation resources simply give out, and the only thing they can do is either shrug their shoulders or, in an attempt to realign them with reality, force their own views onto them.

If Jonesian mindreading doesn't endow the Ryleans with the capacity to attribute *thoughts* as articulating someone's subjective take on the world, then it doesn't enable them to ascribe *beliefs* either. The Myth of Jones in EPM remains silent on matters concerning *dispositional* states of belief. However, in 'Language as thought and as Communication' (1969/2007; hereafter LTC), Sellars uses his verbal behaviorist framework to give an analysis of belief. Building on the notion of thinking already explicated, he analyses 'Jones believes that-*p*' as "Jones has a settled disposition to think that-*p*, if the question occurs to him whether-*p*, and, indeed, to think-out-loud that-*p*, unless he is in a keeping-his-thoughts-to-himself frame of mind." (LTC, p. 75)[38] Suppose we add this analysis to the curriculum of the Ryleans' mindreading course. Will this enable them to achieve a genuine understanding of beliefs as subjective, possibly false representations of the world? No. The thoughts that figure

38 In a footnote he adds: "The 'if the question occurs to him whether-p' condition can be taken to cover all cases in which, where the alternatives 'p' and 'not p' are relevant to his course of thought, he thinks that-*p*, even if the question whether-*p* is not actually raised."

in the analysis are such that their contents are still determined entirely with reference to the public world. When someone is disposed to think a thought that defies the entry/inference/exit rules of the Rylean language game, she will simply be treated as being disposed to fail thinking a proper thought.[39]

Notice that mere mastery of the subjunctive conditional required for an understanding of the notion of having a disposition, does not suffice here. To have a disposition to a is to actualize a under certain conditions. But as far as the (educated) Ryleans are concerned, these conditions are identified solely with reference to actual or possible states of affairs of the public world as described by their pre-Jonesian language. Their understanding of other people's disposition to think that *p* is confined to their own appreciation – and in their view the *only, legitimate, common* appreciation – of the counterfactual situations that specify the content of *p*. The idea that someone under different circumstances might sincerely claim 'in a keeping-one's-thoughts-to-oneself frame of mind' something *contrary to their public appreciation of the counterfactual situation*, this is something that still eludes the Ryleans, and perhaps even genius Jones himself.

Counterfactual thinking is something the Ryleans could already do before the arrival of Jones. After his mindreading classes, they can also think *about* other people's thinking in counterfactual circumstances, provided that other people's take on these circumstances meshes with their own. Their concept of the standing state of belief shares much of the normative impact but lacks the subjective dimension of our concept of belief. The Jonesian idea of sincerely claiming to oneself that p in various counterfactual situations is like our concept of belief in bestowing a commitment to it being the case that p on the believer, but it cannot make room for appreciation of the fact that such commitments need not be shared by anyone else.

An understanding of being disposed to think modeled on an understanding of being disposed to assert that itself is confined to the public realm cannot yield a mature concept of belief. The same goes for a dispositional analysis of desiring *p* in terms of, roughly, being disposed to *intend* to *p* (and, subsequently, having the disposition to act accordingly) under suitable circumstances (e.g. Sellars 1966).[40] The Ryleans' post-Jonesian conception of an agent's intending-

39 This point is not restricted to Sellars's specific analysis; it applies generally to all dispositionalist and/or functionalist accounts of belief. In Sellars's myth, the dispositions or functional roles that are deemed definitive of beliefs on these accounts would also have to be characterized entirely in terms of the Ryleans' pre-Jonesian language.

40 Cf. Smith's (1994) analysis directly in terms of being disposed to *act* to the effect that *p* under suitable conditions: "According to this alternative conception, desires are states that have a certain functional role. That is, according to this conception, we should think of desiring to j as

to-herself to *p* (and acting accordingly) is entirely driven by their pre-Jonesian understanding of her overt avowal 'I shall (now) *p*' (followed at some future time by an action to the effect that *p*), and will thus be interpreted as a successful case of intending only insofar as it is in conformity with the language exit rules that determine proper use of its overt counterpart in pre-Jonesian public language. Inappropriate or unrealistic intentions and actions that defy the rules of the pre-Jonesian language game will be treated as failed expressions of publically acceptable desires, not also as successful expressions of discrepant personal desires. Attributing an unrealistic intention, for example, requires that the interpreter distinguish between what someone is supposed or allowed to bring about according to one's own normative expectations, and what is represented by that person's intention as the goal of the action, something she genuinely believes to be attainable (otherwise it would not be a case of sincere intending), but which the interpreter himself judges to be out of reach. Such representational understanding of intention goes beyond Jones's teachings. Adding a dispositional dimension to this understanding, in the form of desiring that *p*, makes no difference in this respect.

The pre-Jonesian Ryleans had a conception of goals and reasons for action; goals were conceived as what was explicitly avowed in overt expressions of intention; reasons as what was explicitly stated in order to justify such explicit commitments to act. Importantly, these utterances could only be interpreted as intentional acts insofar as they were in accordance with the rules of their public language game; their verbal behaviorist conception of mind did not allow for an understanding of overt thoughts and intentions that somehow failed to comply with the entry/inference/exit rules that characterized proper (linguistic) conduct. Thus, people could only be judged to intentionally adopt goals and respond to reasons insofar as their answers to questions why were deemed acceptable in light of what one *ought* to say and how one *ought* to act given the circumstances, according to the public rules. The descriptions provided of people's goals and reasons, to the extent that their answers *could* be conceived as voicing goals and reasons, portrayed these goals and reasons as things in the public world (past, present or future): events or states of affairs one was expected to achieve, in the light of other events, states of affairs

having a certain set of dispositions, the disposition to y in conditions C, the disposition to k in conditions C', and so on, where, in order for conditions C and C' to obtain, the subject must have, *inter alia*, certain other desires, and also certain other means-ends beliefs, beliefs concerning j-ing by y-ing, y-ing by k-ing and so on." (p. 113) Notice that for the Ryleans, these 'beliefs' can only be regarded as dispositional *relational states*.

or facts that made their achievement worth pursuing by making certain language exit moves, i.e. by performing certain actions. By accepting the answers given to their questions why, Rylean interpreters were in effect *relating* their fellow agents to their goals and reasons out in the world; to them, *all* goals and reasons that people could adopt and respond to were things that either had featured, did obtain or would come to be realized in the world. Thus, successful interpretation of each other's actions implied the past, present, or future existence of the goals and reasons attributed. This is relational mindreading (cf. chapter 2.4).

Jones only taught the Ryleans how to conceive of other people's goals and reasons as being issued silently by their (dispositions to have) thoughts and intentions. He taught them how to move their game of giving and asking for reasons 'inside', but not how to sever the bonds with the public realm in doing so. Jonesian mindreading therefore did not allow for representational mindreading, it did not allow for the attribution to others of subjective representations of the things they intend to achieve, nor of the things that make these ends and their means worth accomplishing. An agent acting on beliefs and desires incompatible with public evaluation of the world could not be regarded as acting with a goal and for a reason. Such an agent surely *tried*, but, on a Jonesian understanding, simply failed to do so.

3.4 Why the Myth Matters

I have exploited Sellars's Myth of Jones with the purpose of making the notion of relational mindreading more tangible. For two reasons, Sellars's myth has been particularly helpful in this respect. First, the end stage of Sellars's myth is still widely conceived as a basically adequate picture of our actual folk psychological concept of mind, the propositional episodes of 'thinking' and related propositional attitudes in particular. Showing that the end stage of Sellars's myth actually depicts a relational stratum of social cognition will help positioning the technical notion of relational mindreading on the philosophical market. Second, Sellars devised his myth as a case of genuine conceptual bootstrapping, of showing how, as he puts it elsewhere

> the explanatory function of 'inner conceptual episodes' can be construed as resting upon an autonomous proto-psychological framework in which linguistic activity is described, explained and evaluated without reference to the framework of 'mental acts' which it supports. (MFC, p. 82)

If my conclusion regarding the end stage of Sellars's myth holds out, then, to the extent that Sellars has succeeded in providing a non-circular account of the concept of inner thought built on the concept of overt linguistic acts, to that very extent will I have succeeded in characterizing a relational concept of propositional attitudes without presupposing a representational notion of mind.

It is important that we be clear about the status of Verbal Behaviorism in Sellars's philosophy. It primarily serves the *expository* purpose of laying bare some important structural features of the conceptual framework of commonsense psychology. He chooses to treat VB as 'methodologically autonomous', i.e. to "present it in the guise of a claim that thinking at the characteristically human level simply *is* what is described by this framework." (MFC, p. 83; emphasis in original) It serves to clarify certain key issues pertaining to the nature of our commonsense concept of thought, but, importantly, "It is not intended to be an adequate account of thinking; it is indeed radically oversimplified." (*ibid*.) The *methodological* autonomy of the VB conception of intentionality need not be taken as implying the (phylogenetic or ontogenetic) *developmental* claim that this rather coarse-grained 'behavioristic' understanding of speech and action is fully acquired independently and prior to 'mentalistic' forms of interpretation. Neither need it be regarded as implying the *psychological* claim that interpretation in actual folk psychological practice tends to be 'behavioristic' by default. The point is that once we see the conceptual structure of our mature folk psychology as revealed by Verbal Behaviorism, the 'radically oversimplified' developmental and practical claims of this expository framework can be left behind, or should at least be treated separately as crude hypotheses for further empirical research.[41] Verbal Behaviorism need not be interpreted as a layer-cake model of the development or practical use of different interpretation techniques. Rather, it is used as a philosophical tool for exposing the conceptual relations in commonsense psychology, once all the conceptual pieces are in place.

41 Cf. MFC (p. 82): "For even if, as I do, one finds a reference to 'inner conceptual episodes' which are only in an analogical sense 'verbal' to be an indispensable feature of what might be called fine-grained psychological explanations, it is nevertheless possible to construe this 'fine-grained' framework as a theoretical enrichment of a 'coarse-grained' behavioristic explanatory framework which, from the former point of view, simply equates thinking with states which are 'verbal' – if I may so put it – in the literal sense. To be interesting for our purposes this 'coarse-grained' framework would have to be methodologically autonomous in the sense hat it would contain categories of sense and reference, meaning and truth which could be fully explicated without any reference to non-verbal 'inner conceptual episodes.' Thus, in this behavioristic framework linguistic episodes would be characterized directly in semantical terms, i.e. without reference to the 'inner conceptual episodes' which, from the standpoint of the enriched framework, are involved in a finer grained explanation of their occurrence."

Sellars's strategy is to develop an alternative account of our folk psychological concepts of inner mental states by treating Verbal Behaviorism *as if* it were an accurate picture of human phylogeny and/or ontogeny. But once the basic functionalist structure of the mental state concepts has been laid bare at the end of Sellars's myth, the behaviorist bootstraps can be cut off, thrown away like Wittgenstein's ladder. In their post-Jonesian frame of mind, internalization of Jones's functionalist theory ensures that the concepts of overt and covert displays of intentionality *mutually* presuppose each other. This is all that Sellars needs to make good on his promise to deliver a viable alternative to the generally Cartesian idea of the Givenness of thought in awareness. For such mutual presupposition suffices to reveal the essentially intersubjective status of the concepts pertaining to such inner episodes as thoughts. The crucial conceptual point of Sellars's proposal is that although such concepts may have 'a reporting use in which one is not drawing inferences from behavioral evidence it nevertheless insists that the fact that overt behavior *is* evidence for these episodes is *built into the very logic of these concepts* …' (EPM 59; emphasis in original) At the end of Sellars's story, all that is required is that we see that we ourselves might very well be using our own folk psychological concepts *as if the myth were true*.

Let me briefly illustrate this point. Lacking the concept of inner episodes of thinking, intending, etc., the pre-Jonesian Ryleans were not able to interpret each other as *using* speech for their own 'covert' ends. They cannot interpret each other's linguistic utterances as *speech acts,* as the term is often used in the literature; speaker intentions of the familiar Gricean sort, for example, are beyond their grasp. On the VB conception, 'thinking-out-loud' is "to be equated with 'candidly and spontaneously uttering "p"' where the person […] who utters 'p' is doing so as one who knows the language to which 'p' belongs." (LTC, p. 68). To know a language is, at a minimum, to use expressions belonging to the language in conformity with the rules of the language (see footnote 7). Making an utterance 'candidly and spontaneously' is an instance of, as we might put it, 'unreflectively speaking one's mind' or, indeed, simply 'thinking-out-loud', that is: appropriately voicing one's language without using it with communicative intentions that are directed at the mental states of others.[42] VB does allow for a conception of 'intending-out-loud' in the form of

42 Cf. LTC (p. 68): "The phrase 'candidly and spontaneously' is intended to sum up an open-ended set of conditions without which the suggestion cannot get off the ground. Jones's thinking that-p obviously cannot be a quoting of 'p' or uttering it on the stage in the course of acting. The qualifying phrase also clearly rules out the case where Jones is lying, i.e. using words to deceive. Somewhat less obviously it is intended to imply that Jones is not choosing his words to express his

e.g. 'I shall now go to the supermarket to by some milk', which is treated as a practical commitment to perform the corresponding action. But such avowals of intentions cannot *themselves* be interpreted *as* intentional actions directed at the 'inner' mind of an audience.

Again, however, this VB treatment of linguistic interpretation need not be regarded as an empirical claim about (the development of) the psychology of actual adult linguistic understanding. In fact, it seems compatible with a Gricean account of the interpretation practice of the *educated* Ryleans. In their educated state, the Ryleans may have evolved into occasional or regular Griceans, treating each other's overt sayings as being caused by speaker intentions, the contents of which include reference to the mental states of the interpreter herself. A Gricean picture of Human communication seems consistent with a broadly functionalist conception of speaker intentions along Sellarsian lines, and, I might add in the present context, with a fundamentally *relational* understanding of such intentions. For of course, post-Jonesian understanding would be confined to a relational interpretation of each other's speaker intentions. The educated Ryleans would not comprehend the idea that others might use their words to deceive them, for example, to willfully make them misrepresent the world, nor would they be able to understand themselves as such deceivers.

These same considerations apply to my exploitation of Sellars's myth. In effect, I have used Sellars's conceptual strategy for my own expository purposes. My aim in this chapter is a purely conceptual one, namely that of making a start at revealing the conceptual coherence of the idea of relational mindreading. At this point, I do not wish to make empirical claims about relational mindreading in actual folk psychological practice. That will have to wait until chapter 5.

Yet as to the matter of conceptual coherence, it may seem that there are some further considerations that pose problems for my characterization of relational mindreading. These concern my use of the term 'relational'. In chapter 2.4, I gave a provisional definition of relational mindreading as implying the past, present or future existence of the items the agent is interpreted as responding to or being directed at. As indicated there, this implication should be understood *from the point of view of the participants*, i.e. interpreter and interpretee. If it is a feature of their common worldview that, say, rats are devils in disguise, then the interpreter may interpret the interpretee's rats-directed

convictions. He is neither lying nor speaking truthfully. In a sense, as we shall see, he is not using the words at all."

intentions and actions by *relating* her to what they commonly presuppose rats to be, viz. devils in disguise, even if, from our point of view, there are no such things as devils in disguise. What matters for our purposes is the *psychology of interpretation*, whether the interpretative act proceeds by ascribing representational or rather relational mental states. Normative questions regarding the ontological commitments exhibited are not of our primary concern. Thus when I speak of relational mindreading and the attribution of relational mental states, it is not my intention to make claims about the ontology implied by the interpretation process described. Our Rylean ancestors may have had ridiculous ideas about the world around them, yet it was to entities of their ridiculous ontology that they, *psychologically speaking*, related each other when making sense of each other's goals and reasons. That, at least, is what the previous sections revealed.

Still, this provisional characterization of relational mindreading seems inadequate in light of our present discussion. As already noted, Jones could have exploited the Ryleans' pre-Jonesian capacity for counterfactual thinking to add a dispositional dimension to the Jonesian concept of thought as inner episode. Such relational concept of belief would then amount to having the disposition to think certain inner thoughts under certain suitable circumstances. Jones, in other words, could easily have taught them to think about other people's thinking *in* counterfactual circumstances. But given their pre-existent ability to engage in counterfactual thinking, we should not be surprised if they also learned how to think *about* their compatriots' thinking about counterfactual circumstances, i.e. to interpret their fellow men and women as making suppositions or considering imaginative scenarios. To carry on where we left off in the previous sections, interpretation of other people's thinking as such on the basis of Jones's teachings could not have yielded a representational understanding of these thoughts. As we have seen, there is nothing in Jones's theory that suggests the idea of inner episodes and related mental states having such representational dimension. But in the case of interpretation of other people's thinking about the counterfactual, the characterization of relational mindreading above does not suffice as the appropriate non-representational alternative.

The problem is that the current characterization does not allow for interpretation of thoughts about things that, as far as the interpreter is concerned, do not exist, have never existed and will never exist. For appropriate interpretation of such thoughts does not imply the past, present or future existence of what it is that is thought about. The notion of relational mindreading needs a slightly more relaxed reading, so as to include these forms of interpretation.

For this we only need to add a disjunctivist clause stating that the things the agent is related ('related') to can *either* be things in the past, present or future world *or* items in a merely supposed context. In both cases, interpretation proceeds by 'relating' another person to the contents of a shared context, *as opposed to* treating her responsiveness *as a (mis)representation of* that context.

For practical purposes, it is vital that the Ryleans manage to distinguish between other people's thinking and talking about the actual and their thinking and talking about the counterfactual – that they know when others are informing them about actual dangers, opportunities, etc. and when they are merely considering what-if scenarios. It is crucial for interpersonal coordination that they understand when someone is merely *entertaining* a thought in making a supposition and when she is *endorsing* a thought as pertaining to the actual world in making a claim.

But this does not require the addition of a representational dimension to their understanding of each other's thinking about the counterfactual. In particular, it does not require seeing the counterfactual context thought about as an *as-if representation of* the real world. In order to think about other people's thoughts about the counterfactual, that is, it is not necessary to treat them as *pretending to believe that* the counterfactual propositions are true when in fact knowing them to be false. Engaging with another person's counterfactual reasoning may simply consist in treating her as *supposing that p* (rather than as *pretending to believe that p)*, when, in fact, *p* (rather than *her believing that p*) is not the case. An invitation, verbal or otherwise (e.g. 'Suppose that…') marks the beginning of engagement with the counterfactual scenario and other behavioral cues mark its limits or end. This enables the interpreter to keep a sharp focus on the distinction between supposition and reality. Explicit commentary determines the context of the counterfactual scenario (e.g. 'Suppose you were to receive one million Euros…') and disagreement about the specifics can be solved by declaration or changing of the rules ('No, you must spend it right away'; 'OK, you can also invest it.') – thought not by arguing about the correct interpretation of the rules.

Thus, the suggested disjunctivist modification of the notion of relational mindreading does not threaten to blur the distinction between the actual and the counterfactual. Mindreading in the post-Jonesian Rylean community is by default strictly relational, bound to the actual world; it is only 'relational' by invitation to engage in counterfactual reasoning. In both cases, however, interpretation is limited to a non-representational understanding of each other's intentional directedness toward the context being talked about. In both cases, that is, interpretation is bound to the public sphere. The idea of thinking *dif-*

ferently ('privately') about the *same* ('public') things, whether hypothetical or actual, has yet to dawn on them.[43]

In the previous section, I argued that on a Rylean (FR-) understanding of thinking(-out-loud), a speaker's claim that *not-p* when in fact *p* can only be interpreted as a failed attempt to assert that *p* and not also as an expression of a false belief that *not-p*. What we have learned from the present discussion, is that Rylean interpreters can *also* understand the speaker's utterance of *not-p* as her *making the supposition that not-p* when in fact *p*. When confronted with a speaker's utterance of a false sentence, the Ryleans can thus *either* interpret her as having failed to indicate something about the *actual* world (past, present or future), *or* regard her as having succeeded in saying something about a mere *counterfactual* scenario. What they *cannot* comprehend is that the speaker considers the 'possible world' the false sentence is about *as actual*. They cannot understand her utterance as an *endorsement* of the thought that *not-p* although in fact *p*, as her *undertaking a commitment* to the effect that it is the case that *not-p*, i.e. as her believing that *not-p*. This requires mastery of a subjective notion of representation, a social skill that supersedes the interpretative capacities of the educated Ryleans.

Although this disjunctivist amendment to the working definition of relational mindreading is necessary to accommodate non-representational interpretation of each other's thoughts about the non-existent, in what follows I will mainly focus on the vast majority of cases in which speech, thought and action are interpreted as responses to the state of the world as it is, as it has been or as it will be. In these cases, the original characterization of relational

43 This discussion runs somewhat parallel to the issue in developmental psychology as to whether the concept of belief is required for engaging in pretend play, which children start doing from approximately two years of age onwards. Following Leslie (1987), Fodor (1992, p. 290) claims that "Pretending involves acting as though one believes that P is true when, in fact, one believes that P is false. It would thus seem impossible for a creature that lacks the concept of a belief *being* false." Following Perner et al. (1994), Harris et al. (1994) and others, Doherty (2009, pp. 95-104) provides an alternative to this characterization, suggesting the following revision of Fodor's claim: 'Pretending involves acting as though P is true when, in fact, P is false. It would thus seem to be impossible for a creature that lacks the concept of a proposition *being* false." (p. 96) This, Doherty argues, adequately captures what is required in pretence, without having to posit metarepresentational understanding. Accordingly, pretence would only require appreciation of the fact that people may act on counterfactual propositions they – the child included – judge to be counterfactual. Understanding the representational state of belief would require differentiating between such cases and instances in which people act on counterfactual propositions they – but *not* the child – evaluate as being actually true. It is currently a hotly debated issue whether children below the age of 4 have a representational understanding of belief (see the appendix to chapter 5 for discussion). If they do, a representational understanding of pretence along the lines suggested by Fodor may be a possibility for them. If they do no not, their understanding of pretence is similar to the Ryleans' understanding of counterfactual reasoning. Putting matters of actual ontogenetic development to one side for now, the present point is merely that the idea of such 'Rylean' pretence seems perfectly coherent.

mindreading still suffices.

But even on the amended definition, the idea of relational mindreading may seem at odds with Sellars's conception of Jonesian thoughts as *inner* episodes. Relational mindreading is supposed to consists in the attribution of relational states, states which relate the 'inner' to the 'outer'. So how can these states also be inner episodes?

The first thing to notice is that Sellars's talk of 'inner' episodes need not be taken to imply such episodes to reside somewhere in the brain or in the head, at least not within the context of the Myth of Jones.[44] As far as the myth goes, Jones may be a rather primitive theorist, he may not know anything about brains, let alone where they are located. Perhaps he believes that thoughts are located in the chest, rather than in the head, or that they run through the whole body. Jones moreover conceives of inner episodes as forming a contrast class with behaviorally manifest 'thinkings-out-loud'. For this contrast, it only matters that inner episodes are not directly manifested in behavior; their location is rather immaterial.

Secondly, and more importantly, even on Jones's theory, an inner episode as an internal or embodied occurrence, when considered by itself, does not constitute a *contentful thought;* it only gains meaning within a wider context. Sellars has Jones model inner thought on over linguistic utterances. With respect to the latter, Sellars argues that the making of a *sound* only becomes meaningful *speech* once it gets incorporated into a behavioral pattern that conforms to the entry/inference/exit rules of the language of the community to which the utterer of the sound belongs (see e.g. LTC; cf. footnote 7). As a result of intensive training, new members of a community start to behave as they should according to the rules. Thus, the causal pattern of the making of a certain sound comes to mirror the rules that specify its meaning as a linguistic utterance. The sound gains its meaning in virtue of the fact that its occurrence is taken up into the socio-culturally shaped causal pattern that ensures that the appropriate transition rules are met, i.e. that the utterance is used in the right entry circumstances, proprietary inferences are drawn and appropriate intentions are formed and actions taken.[45]

Once a member of the linguistic community has reached this stage, interpretation of her utterances as saying something in her native language can

44 Sellars hints at the idea of inner thoughts being realized by brain processes (EMP, §55), but this plays no role in the overall argument of EPM.

45 See especially 'Actions and Events' (1973, especially pp. 192-195). See Rosenberg (2004/2007) and O'Shea (2007, ch. 4) for discussion. Sellars's proposal shows some interesting parallels with Dretske's (1988) idea of 'structuring causes'.

go accompanied by appreciation of the fact *that there are* causal connections between her utterances, the environment and her actions that explain the sequence of events. Thus, the interpreter of a series of meaningful events, e.g. a person's saying "I shall now go to the supermarket" followed by his going to the supermarket, may acknowledge that there is also some causal connection between these events that explains why they occurred. But this is not to say that the interpretation of *what* the person is saying and doing is determined *in terms of* such causal connections. The overt utterance that, qua mere behavioral event, may be considered to be causally relevant in bring about certain other behavior, is also, *qua* meaningful saying, interpreted as being about something in the 'outer' world informing the action to which it has committed its utterer.

These considerations carry over to the attribution of Jonesian inner thoughts. Sellars lets Jones exploit the pre-existent capacities of the Ryleans to engage in semantical and theoretical discourse. Jones teaches the Ryleans how to use the FR-notion of representation in terms of which they were already able to classify their utterances in order to theorize about the causes of such utterances. As mental representations, thoughts are specified in terms of the *functional roles* of the overt utterances they are modeled upon. At the same time, these thoughts also have certain *causal roles* to play in the production of speech and behavior, causal roles, which, from the theoretical point of view, somehow approximate their functional roles as ideally described from the semantical point of view (i.e. in terms of entry/inference/exit transitions). Their causal role as 'inner' episodes in Jones's *explanatory* theory is not to be conflated with their functional or inferential role as contentful thoughts in *interpretation*. If, as has been suggested in the previous section, Jonesian mindreading is a form of relational mindreading, then the Ryleans' conception of inner episodes *as contentful thoughts* portrays them as relational entities. Whereas, *qua* causally relevant 'inner' events, they may be considered not *constitutively* but merely causally related to the outer world, *qua* contentful occurrences, they are interpreted *as being related* to what they are about in the outer world.[46]

46 My treatment of the Rylean's causal understanding of each other's behavior shows parallels with Jackson and Pettit's (1988, 1990) 'program model' of explanations in terms of functional roles. They observe that "we can and often do explain by citing a feature which causally programmes without causing. Features which causally explain need not cause. This is typically what happens when we explain in terms of highly relational properties." (1988, p. 392) The idea is that the psychological properties figuring in folk psychological explanations of behavior are not themselves causally efficacious properties with respect to that behavior, but rather 'program' for or 'ensure' the presence of some other properties (neural properties, say) that are. Morton (2003, ch. 4) suggests that folk psychological explanations are causally 'shallow', essentially contrastive explanations. Likewise, Hutto (2011a) speaks of the *informational* rather than the causal relevance of reasons, pointing out that the fact "that the factors cited by the folk are mention-worthy does not entail that they pick out (or attempt to pick out) causally relevant properties per se." (p. 140)

Jones's explanatory theory of mind only works because it rides piggyback on the pre-existent capacity of the Ryleans to interpret each other's sounds and doings as meaningful sayings and actions (see section 3.5 and appendix for discussion). Some causal story is reflected in the entry/inference/exit transitions that they can comment upon in their functionalist meta-language. This is all that Jones needs to introduce them to the revolutionary idea that this causal story, whatever it is, actually involves silent counterparts of the events they already interpret as meaningful. As contentful thoughts, these 'inner' episodes are interpreted in the same way as the overt utterances they are modeled upon, i.e. as one side of a relation that reaches out into the public world. At the same time, acknowledgment of the fact that these same episodes have some role to play in the causal story makes it possible to think about them as non-relational states in or of the body. But it is not in virtue of such acknowledgment that thoughts are ascribed *as contentful episodes* in the course of interpretation. As far as mindreading is concerned, the thoughts ascribed relate the thinker to those features of the world they are interpreted as being about.[47]

3.5 The Myth of Jones and the Theory Theory

The end stage of Sellars's Myth of Jones has widely been accepted as an accurate picture of our social practice as depicted by the Theory Theory of folk psychology. There has been discussion among theory theorists as to how exactly our folk psychological theory is supposed to be represented in our brains. As indicated in chapter 2, there is disagreement as to whether we should explain our folk psychological capacities in a more nativist, modular way (e.g. Fodor, 1992, Leslie 1994, Baron-Cohen 1995) or rather in a more empirist, non-modular fashion (e.g. Gopnik and Wellman 1994, Gopnik and Meltzoff 1997). Another point of dispute concerns the representational format of the folk psychological

Consider also Schueler's (2003) 'minimalist' treatment of the folk psychological use of the term 'cause', according to which "the claim that one thing caused another amounts to nothing more than saying that the second happened *because* of the first, that is, that the one explains the other, period, with no information at all about how this explanation works." (p. 14; emphasis in original)

47 Thus, the problem of mental causation (cf. Heil and Mele 1993; Kim 1998) is not a problem that Jones or his Rylean students need to solve in order to become fluent Jonesian mindreaders. The problem of mental causation is to explain *why* it is really in virtue of their content that mental states causally explain behavior. This is not something the Ryleans need to be able to explain. All they have to know in order to appreciate Jones's teachings is *that there is* some or other causal story reflected into the sequence of events they interpret as meaningful according to the rules of their language (see footnote 46). In general, successful folk psychological interpretation does not require a solution to the problem of mental causation. The key to solving this problem, if there is such to be found, should not be expected to lie hidden in our folk psychological practice.

theory at the subpersonal level, whether, for example, it is stored sententially in a language of thought (e.g. Fodor 1987) or rather subsymbolically encoded in the distributed weights of connectionist networks (e.g. Churchland 1991, Ramsey et al. 1991). But as far as the *information* conveyed by the theory is concerned, however stored, acquired or applied, all theory theorists have remained loyal to the basic functionalist insight that Sellars developed in EPM.

This is most evident for the empirist, 'scientific' versions of the Theory Theory wedded to tacit but explicitly represented knowledge of the lawlike generalizations linking mental states to input, output and each other (e.g. Gopnik and Meltzoff 1997). Here, our supposedly quasi-scientific folk psychological theory is basically an explicit version of Jones's theory, stating all the nomological regularities between mental states, input and output that Jones's model for inner thought, i.e. Sellars's functional classification of Rylean language use, implicitly conveys. Thus, on these versions of the Theory Theory, the *meaning* of our mental state concepts is given by broadly Sellarsian functional roles in terms of which the theoretical terms of this quasi-scientific theory are defined. While modular theory theorists do not tend to endorse a holistic, functionalist semantics of our core folk psychological concepts (instead opting for e.g. an atomistic, purely denotational account of meaning, cf. Fodor 1990; see Hutto 2008a, ch. 8 for discussion), the total information contained in the theory of mind module(s) must include reference to the causal relations between mental states, input and output if it is to be of any use in explaining or predicting behavior. This is to say that even if our folk psychological mental states concepts are not *defined* in terms of their functional role, the propositions stored in the theory of mind modules must nonetheless contain specifications of such functional roles. And although connectionists hold that our folk psychological theory is not propositionally, sententially represented in our brains, they agree with other theory theorists that propositional articulation in typical functionalist fashion does provide a fairly adequate picture of the conception of mind according to that theory (cf. Churchland 1991; Ramsey et al. 1991).

The tacit knowledge of our alleged folk psychological theory is thus considered to be at least roughly extensionally equivalent to a functional role description of mental states along the lines set out by Sellars in his Myth of Jones. As we have seen, however, Sellars's myth does not give us the conceptual material needed to bootstrap the Ryleans into a representational understanding of mental states. As far as Sellars's myth goes, the Theory Theory of folk psychology only posits *relational* mental states. In this context, consider David Lewis's widely influential account of the semantics of theoretical terms (1970), which he applied to folk psychology in his defining paper on ana-

lytical functionalism (1972). In this paper, Lewis explicitly mentions Sellars's myth as the perfect illustration of his account, proposing to 'adopt the working hypothesis that it is a good myth' (p. 213). It is a good myth, he adds, 'if our names of mental states *do in fact mean just what they would mean if the myth were true*' (*ibid*., emphasis added). Again, however, Sellars's myth only brings us as far as a relational stratum of social cognition. So if, through Lewis's powerful analysis, Sellars's myth is still to be our guide to the semantics of the theoretical terms of our commonsense psychology, then there is every reason to think that some of our propositional attitude concepts, perhaps even the most important ones, take us exactly to the point where Sellars's myth ends: a sophisticated, inferentially articulated, yet *relational* understanding of each other's responsiveness to the world.

This should give us pause. For if the most elaborate functionalist accounts of the semantics of our mental state concepts in fact do not dictate a representational reading of these concepts, then why insist on belief-desire psychology as the single core explanandum of commonsense goal-reason psychology? Since Sellars's myth provides us with nothing beyond a relational conception of mental states, theory theorists would be well advised to reconsider their representationalist assumptions.

This last point should not mislead us into thinking that the conclusions reached thus far are premised on the idea that relational mindreading *is* theoretical in nature. It merely states that *if* folk psychology is indeed a theory, i.e. if our mental state concepts are indeed theoretical terms in an explanatory theory of human behavior, we should seriously consider the possibility that its core theoretical terms refer to relational, rather than representational mental states. Once the idea of relational mindreading is on the table, however, there is nothing that should prevent us from applying similar considerations to other stories about the cognitive underpinnings of mindreading. The current proposal cuts across the different versions of Theory Theory, Simulation Theory, and yet other alternatives. If, as I suggest, interpretation in terms of propositional attitudes indeed falls apart into two forms of understanding, i.e. relational and representational mindreading, then this differentiation of the *explanandum* of folk psychology should be brought to bear on all theories that purport to explain it (see chapter 5.2).

Sellars's *own* proposal in EPM actually falls short of a strictly theoretical rendering of folk psychology. First, determination of the contents of the mental states ascribed in their post-Jonesian stage is based on the Ryleans' pre-Jonesian and non-theoretical linguistic know-how. Jonesian mindreading is parasitic on linguistic skills that are not theoretically mediated on Sellars's

story. Second, mental state ascriptions may be non-inferential social responses on Sellars's account. Many defenders of TT take it to be one of the defining commitments of TT that ascription proceeds by *application* of the theory, i.e. drawing an inference to the best explanation on the basis of behavioral and situational evidence.

As to the first, Jones teaches the Ryleans to exploit their understanding of each other's linguistic utterances for learning his new theory, but there is no suggestion to the effect that the pre-Jonesian capacity the Ryleans exhibited in interpreting each other's sayings is *itself* theory-based. In fact, the way in which Sellars sets up his story points to the contrary. Recall that semantical discourse is presented methodologically as an *addition* to the original Rylean language. The functionalist classification of the use of their language in semantical discourse is parasitic on their first-order skills to appropriately use their language and assess its use by other people. These first-order linguistic skills themselves do not appear to be theoretical on Sellars's story, since the capacity to theorize – the addition of theoretical discourse to their discursive repertoire – is also presented as an enrichment of the original Rylean language. Jones teaches the Ryleans his 'theory' of mind, but there is nothing in Sellars's myth that suggests that prior to Jones's arrival, the Ryleans have internalized a *theory of meaning* that enables them to interpret each other's linguistic utterances. Rather, Jones's theory of mind builds on their pre-existent and, as far as the myth goes, pre-theoretical understanding of linguistic meaning. Once they have become skilled Jonesian mindreaders, the Ryleans determine the *contents* of the mental states they ascribe by exploiting their non-theoretical understanding of the *meaning* of each other's linguistic utterances.[48] Thus, Sellars's Myth of Jones suggests that

> Our attributions of psychological states rest on a kind of pre-theoretic know-how […] they rest on knowing how language and social situations work "from the inside". This knowledge can then be recruited to make sense of both our own and others' behavior without having to take a detour through externalization in an explicit theory of psychology. (deVries 2005, p. 198)

48 The suggestion that content determination rests on non-theoretical linguistic know-how shows obvious parallels with Heal's account of folk psychological interpretation as 'replication' (1986) or 'co-cognition' (1998). Rather than theorizing about the contents of other people's thoughts, we may simply 'replicate' the thoughts of the other, 'co-cognize' with her, so to speak, in order to reach a verdict on the inferences she will draw concerning the subject matter at hand. See chapter 5.3 for further discussion.

Determination of content does not proceed by application of some naturalistic theory of content that specifies the contents of mental states in purely causal terms. As explained in the previous section, it is enough for Jones and his Rylean students to realize that there are some such causal connections reflected in the meaning of their utterances, without having a clue, let alone a theory, as to how these causal relations give rise to content (see appendix for further discussion).[49]

In order to appreciate the second aspect in which Sellars's account does not support a theoretical rendering of folk psychology concerns, we should reconsider Sellars's overall aim in EPM: to provide an alternative for a Given in first-person epistemology. One important point that often goes unappreciated is that the Myth of Jones actually contains two theories: Jones's *psychological* theory that is supposed to explain behavior in terms of mental states and Sellars's *philosophical* theory that aims at elucidating the status and nature of our concepts of mental states (cf. deVries 2005, pp. 178-179; see also Parsell 2010). Jones uses overt verbal behavior as a model for inner episodes; Sellars uses the introduction of theoretical concepts in science as *his* model for the *concepts* of inner episodes. Like Jones's theory, Sellars's theory also contains a commentary that places restrictions on the analogy suggested between the model and the subject matter of the theory. Jones's proposal is that inner thought is like overt speech to the extent that it is semantically evaluable; it does not go accompanied by the making of inner sounds or the wagging of inner tongues. Sellars's proposal is that folk psychological concepts are like theoretical concepts insofar as they have an essentially *intersubjective status*. Aside from a few brief remarks, Sellars does not provide an explicit commentary on the model.[50] Yet the limits of the analogy can be gleaned from his overall strategy in EPM.

The analogy with scientific theorizing is Sellars's antidote for the myth

49 Cf. deVries (2005, p. 198): "They (and we) can assume that there are causal connections among our internal states that enable those internal states to relate to each other in ways that conform to their semantic characterization, but how those causal connections work, what kinds of mechanisms instantiate them, how they might develop and how they might break down, are left untouched. The "internal structures" recognized in Jonesian psychology are almost completely parasitic on the structure of the language Jones and company speak."

50 But consider: "I am going to argue that the distinction between theoretical and observational discourse is involved in the logic of concepts pertaining to inner episodes. I say 'involved in' for it would be paradoxical and, indeed, incorrect, to say that these concepts are theoretical concepts." (EPM, §51) And: "I have suggested a number of times that although it would be most misleading to say that concepts pertaining to thinking are theoretical concepts, yet their status might be illuminated by means of the contrast between theoretical and non-theoretical discourse." (EPM, §59)

of the Given. He contrasts theoretical terms and entities with observational terms and observable entities, respectively. What makes something a theoretical entity is that one can only come to know about it by making inferences on the basis of observable phenomena. Observable entities can also be known non-inferentially; observational terms can be used in making non-inferential *reports*. As Brandom comments: "To be observable is just to be non-inferentially reportable." (1997, p. 164). We might add: to be theoretical is just not to be non-inferentially reportable.[51] Sellars uses this contrast to show how what is *like* a theoretical concept in being essentially intersubjective (and therefore not Given) may come to be used in making non-inferential reports – and hence may actually *seize* to be strictly theoretical – *without* losing its intersubjective status. This is the final stage in his myth, were Jones conditions the Ryleans into first-person *non-inferential* use of his newly introduced concepts of inner episodes: "*what began as a language with a purely theoretical use has gained a reporting role*" (EPM 59; emphasis in original).[52] Sellars here tries to combine the 'theoretical' – i.e. essentially intersubjective – status of the concept of inner episodes with its privileged, non-inferential – i.e. *non-theoretical* – use in first-person ascription.

Sellars does not extend this idea to second- or third-person ascription of mental states in EPM. But there is nothing in Sellars's philosophy against the idea that in a limited range of conditions, interpreters can report non-inferentially on the propositional attitudes of others (cf. deVries 2005, pp. 194-195). On Sellars's story, an experienced doctor, say, may acquire the reliable differential responsive disposition to non-inferentially see that his patient has lung cancer by looking at a chest-X-ray. The less experienced, in contrast, may only see a particular pattern of white-grey clouds from which they then *infer* that it

51 Sellars's distinction between observable and theoretical entities does not parallel the distinction drawn in chapter 2.6 between observable and unobservable mental states. There, the contrast was used to characterize the nature of the target states attributed. Observable states closely track specific behavioral types, whereas unobservable states, due to their holistic nature, only show tenuous connections to behavioral types. Propositional attitudes are typical unobservable states in this sense of the term. Sellars's distinction pertains to the interpretation process, however. Here the contrast is between entities that can be perceived directly and entities that can only be inferred on the basis of other evidence.

52 The text continues: "As I see it, this story helps us understand that concepts pertaining to such inner episodes as thoughts are primarily and essentially *intersubjective,* as intersubjective as the concept of a positron, and that the reporting role of these concepts – the fact that each of us has a privileged access to his thoughts – constitutes a dimension of the use of these concepts which is *built* on and *presupposes* this intersubjective status. […] it also makes clear that this privacy is not an "absolute privacy." For if it recognizes that these concepts have a reporting use in which one is not drawing inferences from behavioral evidence, it nevertheless insists that the fact that overt behavior *is* evidence for these episodes *is built into the very logic of these concepts*, just as the fact that observable behavior of gases is evidence for molecular episodes is built into the very logic of molecule talk."

is probably lung cancer. Similarly, an experienced folk psychological interpreter may acquire the disposition to respond adequately, yet non-inferentially, to the mental states of others. For people less experienced in social life, or for those who do not know this particular person well, ascription of the relevant mental states may only be possible through inference to the best explanation from contextual cues.

In the current debate on social cognition, allowing for non-inferential use of folk psychological concepts in the interpretation of other people would be a major concession for the Theory Theory of folk psychology. Some authors have recently framed the debate in terms of the question whether social understanding should be regarded as a form of non-inferential social perception (e.g. Gallagher 2008b, 2011, Zahavi 2007, Zahavi and Gallagher 2008, Zahavi 2011) or rather as involving inferential mindreading (Herschbach 2008a, 2008b, Spaulding 2010). Many defenders of TT take it to be one of the defining commitments of TT that ascription proceeds by *application* of the theory, i.e. drawing an inference to the best explanation on the basis of behavioral and situational evidence. Thus, Herschbach (2008b, p. 223) represents the majority of theory theorists when he says: "Mental state attribution occurs for Theory Theory via theoretical inference, by applying theoretical knowledge about the relations between observable behavior, environmental context and mental states."[53] In light of the current dialectics, the suggestion of non-inferential use of mental state concepts in social perceptual reports would firmly place Sellars in the anti-Theory Theory camp.[54]

Having said this, it is important to realize that while, on Sellars's story, second- and third-person mental state ascription may thus proceed non-inferentially (and hence non-theoretically), for it to be treated as a genuine case of ascription, i.e. as a report with propositional content, this *non-inferential* response needs to have an *inferential role*. What makes the experienced inter-

53 Cf. Spaulding (2010, p. 121): "Theory Theorists argue that we attribute and theorize about mental states by employing folk psychological theories about how mental states inform behavior. With our folk psychological theories, we infer from another person's behavior what his or her mental states probably are. And from these inferred mental states, plus the psychological laws in the theory connecting mental states to behavior, we predict the behavior of the other person."

54 Notice that the issue whether or not propositional attitudes are ascribed through inferential processes runs *orthogonal* to the issue whether the attitudes ascribed are conceived as relational states or as genuine representational states (beliefs, desires). Perhaps propositional attitudes can be directly perceived in the behavior of others in relatively familiar circumstances. It may be the case that I am able to see the thought in my friend's face that I have heard him express so many times before in similar situations. But even if one has to do a lot of conscious reasoning in order to figure out why someone acted the way she did, it is perfectly possible that this reasoning involves reference only to *relational* attitudes – drawing inferences about the world the other person is conceived as having access to.

preter's non-inferential ascription a report to the effect that e.g. his neighbor bought a new car in order to impress his neighbors, is, *inter alia*, its specific pattern of inference: the claims one is committed and (not) entitled to draw from it or to draw it from, etc. We thus need to distinguish 1) the *manner* of ascription from 2) the *significance* of the ascription. While 1) may proceed both inferentially and non-inferentially on Sellars's account, 2) must be *inferentially articulated* in order to be meaningful in the first place and play a proper role in the game of giving and asking for reasons. This is particularly relevant for the ascription of states with propositional content in providing goals and reasons for action. While such ascription may thus be reached non-inferentially, the states ascribed must have an inferential role.[55]

We should conclude that the idea of relational mindreading, as illustrated in this chapter by means of Sellars's Myth of Jones, is not committed to a specifically theoretical rendering of the interpretation processes involved.

3.6 Conclusion

This chapter has been an exercise in 'dishabituation': an attempt to counter the habit of talking about goal-reason psychology in terms of representational belief-desire psychology. Sellars's Myth of Jones has been a particularly useful tool in this respect. Sellars presents his myth as a means of conceptual bootstrapping, of showing how one can bake a truly mentalistic cake out of respectable 'Rylean' ingredients. I simply exploited Sellars's strategy: the functionalist conception of mind that one gets by merely pulling the Rylean conception inside is a relational conception, not a representational one. Following Sellars's verbal behaviorist methodology, this chapter has thus presented us with a way of reflecting on goal-reason psychology without committing the reflective fallacy.

Now that we have bootstrapped ourselves into a relational conception of the propositional attitudes, however, we can go beyond Sellars's particular philosophical framework. The next chapters will put the notion of relational mindreading to the test, by plugging it into other theories of goal-reason attribution. As we shall see, a relational conception of mindreading is not wedded to any specific account of the psychology of goal-reason psychology. This,

55 The inferential roles of propositional attitudes mirror their status as unobservable states in the sense used in chapter 2.6 (see footnote 51). Non-inferential ascription of propositional attitudes yields them as observable states in Sellars's sense of the term.

in turn, will clearly reveal what has already been suggested in this chapter: that the absence of a relational conception of the target explanandum of folk psychology should be regarded as a serious lacuna in the philosophical and psychological study of human social cognition.

Appendix: Relational Mindreading and Functional Role Semantics

The pre-existent Rylean meta-language that Jones exploits for his teachings gives expression to what Sellars later developed into his functional role account of semantics. The point of Sellars's functional role semantics, however, is to give a non-relational account of meaning statements and of corresponding statements about the intentional contents of mental states. This may give rise to the worry that something must have gone wrong in my argument in the main text of this chapter. For how can a theory of mind framed in terms of what is meant as providing a non-relational account of the meaning of linguistic acts, itself render a relational understanding of mental states? This worry will be addressed here.

For reasons that go well beyond the scope of this chapter, Sellars finds a relational conception of meaning unsatisfactory. On Sellars's functional role conception of meaning, meaning statements of the form "'*x*' (in L) means *y*" do not give expression to there existing a relation of some sort between a linguistic item on the left hand side and a non-linguistic entity on the right hand side. In this respect, Sellars's semantics contrasts sharply with other popular accounts, such as causal theories of meaning or the different varieties of truth-conditional semantics. Sellars does not deny that there must obtain relations between linguistic items and worldly entities for the former to have empirical meaning. For 'red' to mean what it does, it is vital that the expression is regularly used in statements made in perceptual response to red things, for example. In general, there must be certain reliable, causal relations between linguistic expressions and things in the world in order for the expressions to have empirical meaning. But from this it should not be concluded, according to Sellars, that meaning statements *themselves* function to refer to a relation of any kind between a linguistic item and a non-linguistic entity. As O'Shea (2007) puts it:

> Rather, the role of 'means' might be such that the truth of a meaning statement entails that there must be certain kinds of empirical-causal relations established between persons' utterances (and thoughts) and various entities, without there being any such thing in the world as a philosophically problematic *meaning relation* holding between those utterances (and thoughts) and those entities. (p. 56)

According to Sellars's analysis of meaning statements, the causal relations

between linguistic utterances and the world are indirectly reflected in the descriptions of the functional roles of those utterances.

On Sellars's account, there are three types of norm-governed linguistic-behavioral patterns that should be manifested for anything to count as a language: (i) language entry transitions (from world to language: perception), (ii) intra-linguistic transitions (from language to language: inference) and (iii) language departure transitions (from language to world: action). In general, as a speaker of the language, one ought to, *inter alia* and *ceteris paribus,* (i) make certain utterances in perceptual response to certain objects or states of affairs (and not make certain other utterances), (ii) be disposed to make certain inferences on the basis of one's own or other people's statements (while refraining from making certain other inferences), and (iii) respond to certain utterances of the form "I shall now a" by displaying certain kinds of behavior (but not by displaying certain other kinds of behavior). Thus, to be a competent user of the English word 'red', for example, one ought to, *inter alia* and *ceteris paribus*, (i) reliably respond to the presence of red objects by uttering 'this is red' (but not by uttering e.g. 'this is green'); (ii) be disposed to make certain inferences, e.g. from 'this is red' to 'this is colored' (but not to e.g. 'this is green'); (iii) reliably respond to one's own utterances of e.g. "I shall now move the red object" by moving the red object (but not by moving e.g. the green object).

A full characterization of the functional role of an utterance type in a language would encompass a detailed description of all the language entry, intra-linguistic and language departure transitions that are appropriate for a user of tokens of that type in that language. The entry and departure transitions describe relations between linguistic items and worldly entities; but these are ordinary causal relations, not basic semantical or intentional relations which meaning statements are supposed to assert. Spelling out the appropriate use of an expression in our language in terms of explicit rules is something 'which we would find difficult if not (practically) impossible' to do (MFC, p. 96) "In practice," therefore, "the use of meaning statements is indispensable, for it provides a way of mobilizing our linguistic intuitions to classify expressions in terms of [their] functions…" (ibid.) Sellars proposes to treat meaning statements of the form "'*x*' (in L) means *y*" as *illustrating* for the speaker (and hearer) of such statements the function of the expression on the left hand side by *assimilating* it to the function in her own language of the expression on the right hand side. Since the utterer is a competent speaker of her own language, she can 'mobilize her linguistic intuitions' regarding the appropriate use of the expression on the right hand side in stating the functional similarity with the expression on the left hand side.

On Sellars's analysis, then, "'*x*' (in L) means *y*" roughly amounts to "tokens of the expression type 'x' in L *have the same functional role* in L as tokens of the expression type 'y' in our language." In Sellars's notation: "'x's (in L) are •y•s". Here a '•y•' is a sortal term, applying to any item in any language that plays that role, viz. the role 'y's play in the home language. The dot notation indicates that the expression on the right hand side is mentioned in a special way, i.e. as *illustrating* the function of tokens of that expression type. The utterer of the meaning statement does not *use* the expression on the right hand side; she rather *talks about its use* in her language, stating that the expression token on the left hand side falls under the functional sortal illustrated by the dot quotes. Thus "'rot' (in German) means *red*" on Sellars's analysis becomes "'rot's (in German) are •red•s", which asserts that tokens of the type 'rot' have the same (similar) functional role in German as tokens of the type 'red' in English for English speakers. In Sellars's own words: "To say that ''rot' means *red*' is not to describe 'rot' as standing "in the meaning relation" to an entity red; it is to use a recognized device (the semantical language game) for bringing home to a *user* of 'red' how Germans use 'rot'." (Some Reflections of a language Game, 1954/2007, p. 39, emphasis in original) On this analysis, "*meaning is not a relation* for the very simple reason that 'means' is a *specialized form of the copula*." (MFC, p. 95) Formally, that is, a meaning statement of the form "'*x*' in L means *y*" does not read 'aRb', expressing a relation R between the linguistic particular on the left hand side and something referred by *using* the expression on the right hand side. Rather it is of the general sortal form 'a is an F', or, simply, 'Gs are Fs', stating *inter-linguistic* functional equivalence (similarity) of two expression types (cf. O'Shea 2007, pp. 55-63; see also deVries 2005, chapter 2).

Let us now return to the issue at hand: is Sellars's non-relational account of meaning compatible with a relational account of Jonesian mindreading? As a first approximation, it should be noted that Sellars's primary target here is *semantical* discourse, i.e. to give an account of the meaning of meaning statements. His non-relational account of 'means' in meaning statements does not automatically carry over to folk psychology, which is primarily about interpretation of *non*-semantical, first-order thought and talk.

The dot quotes in Sellars's analysis of meaning statements serve to indicate a functional classifier that *illustrates* for the *user* of the right hand side expression the function of the expression on the left hand side, drawing upon the user's 'intuitive' knowledge, as Sellars calls it, regarding appropriate use of the right hand side expression. This know-how does not only concern the speaker's own use of the expression but also her assessment of its use by oth-

ers, her ability to judge which language entry, intra-linguistic and language departure transitions are appropriate for another speaker. In other words, Sellars's non-relational meaning statements rely, *inter alia*, on second- and third-person interpretation skills of the user of the language for whom the dot-quoted expression is supposed to serve as an illustrating functional classifier. But these interpretation skills appear to be relational on Sellars's story. That is: assessment of the appropriateness of the use of an expression by others seems to proceed by, inter alia, relating others' perceptual statements to the entry conditions that evoke them and relating their avowals of intention to the exit conditions that satisfy them. In any case, this is how things must work for the Ryleans.

In order to see this, it should be realized that Jones taught the Ryleans a theory of mind by *exploiting* their linguistic know-how. He did not teach them a *theory of language*. As argued in section 3.5, there is no reason to think that the Ryleans' interpretation of each other's overt verbal behavior *itself* depended on a functionalist theory of language. To suppose that it did is to regard pre-Jonesian interpretation to consist in the application of a theory that explicitly states the transition rules that specify the functional roles of the expression in their language. There is no support for this claim in Sellars's myth. As far as the Ryleans are concerned, their functional role semantics is not a theory that *explains* the meaningfulness of their utterances; it is only a device that *classifies* their meaningful utterances in a functionalist way, so as to enable them to comment upon and criticize each other's linguistic performances. These functional classifications are parasitic on a prior, non-theoretical understanding of each other's verbal behavior. It is this classificatory apparatus that Jones uses to characterize the contents of the theoretical posits of his new theory, the inner episodes he calls 'thoughts'.

In order to determine the meaning of a linguistic utterance or the content of a mental state there seem to be, in essential outline, two options. Either one draws upon one's own recognition, emotion and response mechanisms (though not conceived as such) in order to determine the worldly features talked or thought about by the interpretee, or one applies a theory that explains not only how an utterance or thought can be about something (i.e. explains intentionality) but also what a specific utterance or thought is about (i.e. determines meaning or content). Importantly, for this second option to be a genuine alternative the first, it should not draw upon the applier's own recognition, emotion and response mechanisms for its explanatory purposes. This means that the worldly conditions featuring in the theory (as entry and exit conditions, for example) should not be coined in commonsense terms,

the use of which would depend on the interpreter's own recognition, emotion and response mechanisms. Rather they should be stated in properly naturalistic terms. This, in effect, is precisely what naturalistic theories of content have been aiming for in the past decades. It is, to put it mildly, highly contentious to argue that folk psychological interpretation depends on a tacitly applied naturalistic theory of content (see also chapter 5.3). If we shy away from this bold claim, however, we should also discard the second option of explaining the interpretation (determination) of meaning and content. The first option reveals that, in one important respect at least, pre-Jonesian interpretation of linguistic utterances was non-theoretical. It also shows in what sense pre-Jonesian interpretation was relational: in interpreting their fellows' utterances, they related them to the commonsense worldly features these utterances were (supposed to be) about.

By modeling thought on speech, Jones teaches the Ryleans to exploit their relational linguistic know-how for the purposes of mindreading. At first instance, this demands use of the pre-existent Rylean functionalist meta-language. Jones shows them how the thoughts that cause their overt utterances can be characterized in terms of the functional classifications of the overt utterances themselves. Thus, for example, he teaches them that an utterance in their language of the sort 'It is raining!' is caused by an inner episode of thinking an •It is raining!•. As the Ryleans start practicing, they at first explicitly go through the different steps of Jones's teachings. Thus, when they hear someone say 'It is raining', they first think to themselves "This is an •It is raining! • caused by an inner •It is raining! •". And when they see someone carrying her umbrella upon leaving the house, they might infer that her behavior is caused by an •It is raining! •, even though it did not result in an overt utterance 'It is raining'. But there is no reason to assume that the educated Ryleans will remain bound to characterization of the thoughts of others in terms of their functionalist meta-language.

Consider learning a second language. On Sellars's story, one would first use translation rules such as "'Es regnet's are •It is raining•s". But learning to speak a second language *fluently*, one presumably starts talking and thinking in that language *itself* at one point, rather than applying translation rules such as the above to every single instance of using the language or interpreting the utterances of others. Similar considerations apply to Sellars's myth. Thus, the Ryleans may start to interpret each other's silent behavior *directly* in terms of 'inner speech', rather than indirectly by functional classification of the thoughts attributed. Their training may reach a point, that is, where they are *as fluent* in reading each other's minds as they are in interpreting each other's

linguistic utterances. At this level of sophistication, mindreading proceeds by 'hearing' other people say things to themselves when going about their business, things the meaning of which is determined in the same relational fashion as pre-Jonesian interpretation of overt linguistic utterances.

Thus, there appears to be no reason why pre-Jonesian interpretation of ordinary discourse and Jonesian mindreading of analogous thoughts could not be relational on Sellars's story, and in fact some good reasons why it should be. It is crucial for Sellars's proposal that the functional roles of linguistic expressions are *illustrated* for the *user* of that language in making meaning statements. When combined with Sellars's expository story of Jones and his Ryleans, this strongly suggests a relational understanding of first-order language use. In certain important respects, then, the plausibility of Sellars's non-relational account of the meaning of meaning statements rests on assumptions pointing in the direction of relational linguistic interpretation.

John McDowell's critique on Sellars's non-relational account (1998/2009) seems to be at odds with this conclusion. In earlier work, McDowell famously argued against a 'sideways-on' picture of the relation between mind and world, a picture according to which the mediating role of experience between our empirical judgments and the world is a purely causal one (cf. McDowell, 1994). This mere causal rendering of the role of experience is hopeless, McDowell argues, "at least as a picture of how things are from the standpoint of experience" (1994, p. 51), that is, as a transcendental, Kantian story of how our empirical judgments come to have empirical content and appear to us in experience as rationally constrained by the world itself.

Our concerns are not primarily transcendental in nature, nor do they touch upon the role of experience in establishing the relation between mind and world. But if we look at McDowell's position from the point of view of (second- or third-person) folk psychological interpretation, his alternative to the 'sideways-on' picture comes very close to the idea that interpretation proceeds by relating each other's utterances and actions to the commonsense world, that the relations drawn between their minds and the world, moreover, are never *merely* causal but also always intrinsically contentful, i.e. contentful in a way that is not analyzable in purely causal terms.[56] In order to interpret

56 The link between McDowell's more transcendental concerns and the nature of folk psychological interpretation is suggested by McDowell himself when he discusses Davidson's radical interpretation (1994, pp. 34-35): "What I do mean to rule out is this idea: that, when we work at making someone else intelligible, we exploit relations we can already discern between the world and something already in view as a system of concepts within which the other person thinks; so that as we come to fathom the content of the initially opaque conceptual capacities that are operative within the system, we are filling in the detail in a sideways-on picture – here the conceptual

others as 'having the world in view' (McDowell 1998/2009), we have to relate them to the world *as it is commonsensically conceived*. But in doing so we rely on our own 'second nature', as McDowell (1994) calls it, our own evolutionarily and socio-culturally shaped cognitive and emotional capacities for carving up the world.

I have suggested above that such folk psychological conception of 'intentionality as a relation' (cf. McDowell 2009, ch. 3) is compatible with Sellarsian functional role semantics. McDowell, however, criticizes Sellars's account precisely for precluding a relational understanding of propositional, conceptually contentful utterances and mental states. He construes Sellars's functional role semantics as an attempt to *constitutively explain* the meaningfulness of utterances and the contentfulness of mental states in terms of the rule-governed uniformities of entry/inference/exit transitions that determine the functional roles of the relevant utterances and states. Some of Sellars's remarks indeed point in that direction. As we have seen, Sellars regards illustrating meaning statements of the form "'x's (in L) are •y•s" indispensable in practice, because it is practically impossible to explicitly state all the rules that govern the patterns in appropriate language use. Nevertheless, he thinks that "The rule governed uniformities […] which constitute a language (including our own) can, *in principle*, be exhaustively described without the use of meaning statements." (1980, p. 92)[57] McDowell takes this remark to reveal Sellars's aspiration to constitutively explain the normativity of meaning "from outside the semantical" (2009, p. 61), i.e. from a sideways-on, non-participatory point of view.[58]

Although this may very well have been one of Sellars's bolder philosophical ambitions, I hope to have shown that it does not follow from his non-

system, there the world – that has been available all along, though at first only in essential outline. It must be an illusion to suppose that this fits the work of interpretation we need in order to come to understand some people, or that a version of it fits the way we acquire a capacity to understand other speakers of our own language in ordinary upbringing."

57 This quote is from 'Meaning and ontology', chapter 4 in 'Naturalism and Ontology' (1980). Interestingly, the original version (MFC, 1974/2007) doesn't contain the clause 'in principle' as in the quote above.

58 Cf. McDowell (2009, pp. 60-61): "On Sellars's interpretation, the content of a statement of significance is a reflection, into a statement of a relation within the conceptual order, of relations that there ought to be, according to the proprieties that constitute a linguistic practice, between two sets of elements in the real order, one of which comprises linguistic items considered in abstraction from the practical proprieties in virtue of which they are meaningful at all. The "ought" with which meaning and aboutness are fraught gets into the picture as a sentential operator, in whose scope there occur specifications of relations that would ideally hold between linguistic items so considered and other elements in the real order. The content of the "ought" with which some fact about significance is fraught – what it is that, according to the "ought" in question, ought to be the case – can be factored out from the statement of significance and specified in terms that are not themselves meaning-involving."

relational account of meaning statements, when considered by itself. Quite to the contrary, the psychological plausibility of Sellars's account of meaning statements in human linguistic practice hinges on the *illustrative* role of the dot-quoted expressions that figure on the right-hand side of meaning statements. This illustrative role may very well build on a relational interpretation of each other as first-order language users, an understanding according to which the aboutness of linguistic utterances and the directedness of actions are already treated as intrinsically meaningful and contentful.

McDowell contrasts Sellars's functional role semantics with Tarskian/Davidsonian truth conditional semantics. On Davidson's account (e.g. 1967/2001b), meaning statements of the form "'x' in L means y" are transformed into "'x' in L is true if and only if y". Crucially, the expression on the right hand side is used by the utterer of the meaning statement to refer to the conditions that make true the mentioned expression on the left hand side, thereby in effect relating that expression to its truth conditions. On this account, McDowell argues, "we relate the conceptual order to the real order, mentioning elements of the real order by making ordinary uses of the words on the right-hand sides of these statements. But we affirm these relations without moving outside the conceptual order – without doing more that employing our conceptual capacities." (2009, p. 63) I agree with McDowell insofar as he holds that a Davidsonian rendering of meaning statements is perfectly compatible with a relational account of interpreting each other as having the commonsense world in full view (although the former does not seem to imply the latter). It may even suggest itself more readily than the Sellarsian account. But in order to decide between the two, other philosophical considerations should be brought to bear that go beyond the issues with which we are presently concerned.

4

Relational, Representational, Subjective

4.1 Introduction

Commonsense explanations of people's actions are often framed in terms of the goals with and the reasons for which they act. Our question concerns the underlying structure of such explanations. The BD-Model claims that our commonsense understanding of others as rational agents evolves around the concepts of belief and desire. Accordingly, interpreting someone as adopting goals in the light of reasons hinges on the ascription of desires representing those goals and beliefs representing those reasons. On the Relational Model, by contrast, our primary epistemic route to the practical concerns of others is essentially world-bound. It is a form of what I have termed 'relational mindreading' and consists in the attribution of states relating people to their goals and reasons out in the public world.

Chapter 3 enabled us to isolate a relational stratum of goal-reason psychology in Sellars's Myth of Jones. The aim of this chapter is to show that the basic idea of relational mindreading does not depend on any particular conception of the propositional attitudes attributed. Section 2 provides a short review of other accounts that have been challenging the BD-Model of folk psychology: Ratcliffe's (2007) and Perner's (e.g. 1991) respective 'situational'

approaches and Gordon's (e.g. 2001) 'factive' account of reason explanation. These accounts also seem to point in the direction of relational mindreading, but, interestingly, object to a Jonesian rendering of the attitudes attributed in terms of FR-representational states. In order to accommodate these accounts, I propose a distinction in section 3 between 'first-order' and 'second-order' relational states, and, correspondingly, between first- and second-order relational mindreading. FR-representational states are typical second-order states; first-order states lack specifications in terms of functional role.

Section 4 then asks the critical question whether the relational constraints on Jonesian mindreading are explained by the specific notion of mental representation that Jones used to teach the Ryleans his theory: the functional role notion. Is there some other account of mental representation that would have enabled Jones to bootstrap the Ryleans into genuine belief-desire psychology? The answer to this question will turn out negative. Exploiting Sellars's expository story once more, we will come to see that the social impairments of the Ryleans cannot be alleviated by merely introducing them to some or other concept of mental representation. Genuine understanding of belief and desire involves appreciation of the fact that how others represent the world may defy the public norms of representation; in addition to merely attributing mental representations to others, it requires the capacity to incorporate information *incompatible* with the public view on the world into the content clauses of the world-directed representational attitudes ascribed. Belief-desire psychology adds an essentially subjective or private dimension to a second-order relational understanding of mind. We can thus distinguish between the attribution of mental representations *simpliciter* as a means of relational mindreading and the attribution of *subjective* representational states as required for genuine belief-desire ascription.

With these distinctions in place, I argue, we have enough material to meet the first challenge laid out for Relational Model in chapter 1: that of meeting the minimal demand of conceptual validity of the notion of relational propositional attitudes and the idea of relational mindreading.

4.2 Situational Understanding and Factive Explanation

Relational mindreading consists in the attribution of relational mental states, states that relate the agent to his goals and his reasons out in the public world as it presents itself to the interpreter: events or states of affairs that the agent is expected to accomplish in the light of events, states of affairs or facts that

make his action an appropriate or the right thing to do. Ratcliffe (2007) seems to have a similar idea in mind when he observes that "the reasons people offer for actions often take the form of simple assertions about features of a situation." (p. 97) He directs attention to short question-answer dialogues, such as (Q) "Why did she turn left?" (A) "The road to the right is one-way." Or (Q) "Why is he in a hurry?" (A) "His bus is about to leave." (*ibid.*) According to Ratcliffe, however, what examples such as these reveal is that "What often is expected is a description of the situation that makes clear the relevant norms of activity, *rather than an account of people's psychological predicaments*." (ibid., emphasis added) Ratcliffe claims, correctly in my view, that folk psychology and explanations in situational terms have a similar structure. He continues:

> Just as one can say 'if B believes p and desires q, all things being equal, B ought to do r', one can say 'if p is the case and q is the case, all things being equal, B ought to do r'. Norms are integral to the relationships that comprise situations, just as many proponents of FP claim that they are integral to the relationships between beliefs, desires and actions. The systematic structure we require in order to interpret people is out there in the shared world. So the burden need not be carried by a complicated understanding of the relationships between mental states. (pp. 97-98)

As I have characterized relational mindreading, the systematic structure we find in the shared world between goals and reasons is what we relate the agent to when interpreting him in terms of his goals and reasons. We interpret him as *intentionally directed at* his goals and his reasons, perhaps as reasoning about which goal to adopt in the light of which reasons. On the current proposal, our understanding of the systematic structure between the agent's goals and reasons simply *is* an understanding of the relationship between his mental states – mental states that relate him to his goals and reasons in the shared world. Ratcliffe's critique is directed at folk psychology conceived as belief-desire psychology (see also Ratcliffe 2006, 2009). Of course I concur that attributing goals and reasons does not require the ascription of corresponding beliefs and desires, considered as representational mental states. But by saying that we often make sense of people's reasons "by referring to aspects of situations, rather than to psychological states" (2007, p. 186), Ratcliffe runs the risk of throwing out the baby with the bathwater. Understanding people in terms of the norms embedded in the shared world, norms that tell us what in a given situation serves as a reason to perform a certain action, can at the same time be a form of truly mentalistic interpretation in terms of relational propositional

attitudes.

Much depends on how precisely we are to understand 'relational propositional attitudes'. What Ratcliffe appears to be objecting to, is the view that understanding people's actions in terms of reasons requires that we ascribe 'internal' mental states to them (*ibid.*, p. 186), states that represent the situation at hand and cause the ensuing action. Our discussion of Jonesian mindreading in the previous chapter suggests an understanding of relational states that corresponds to this view. Recall that according to Jones's theory, overt behavior is caused by inner episodes whose causal roles mirror the functional roles (in terms of specifiable entry/inference/exit rules) of the overt utterances that either make up the behavior to be explained or would have made most sense in the context of the behavior to be explained. As we have seen, the semantical characterization of inner episodes in terms of functional roles amounts to an FR-representational understanding of such episodes. What Jones taught our Rylean ancestors, then, is how to causally explain each other's behavior with reference to FR-representational states, states to be interpreted analogously to the overt utterances on which they are modeled.

But this is not the only way to conceptualize relational propositional attitudes. Consider Perner's (1988, 1991) ontogenetic account of folk psychology, for example. Interestingly, Perner (1988) *contrasts* propositional attitudes with mental representations. He uses the term 'propositional attitude' to refer to non-representational 'situational attitudes': attitudes that relate an agent to a situation in propositionally articulated ways. Perner argues that children under the age of 4 are mere 'propositional attitude theorists' (1988) or 'situation theorists' (1991). As situation theorists, 3-year-olds can understand that other people evaluate descriptions of a situation as true or false. But what they cannot understand, according to Perner, is that their propositional attitudes towards a situation are mediated by mental representations. This requires a 'representational theory of mind', which, Perner argues, children acquire around the age of 4.[59] This 'theory change' should not be understood as a

59 Perner's notion of propositional or situational attitude shows some similarities with what Flavell (e.g. 1988) terms 'cognitive connections'. On Flavell's account, children of 2-3 years old have learned that other people can be 'cognitively connected' to things in the world in a variety of different ways. They understand, for example, that one may become cognitively connected to something by means of different sense-modalities (e.g. seeing it or hearing it) and attitudes (e.g. thinking about it or remembering it), that these connections can change over time, that their own connections to external objects are independent of those of other people and that they go accompanied by inner experience. However, "young children tend not to understand that forming cognitive connections to things entails mentally representing those things in various ways." (1988, p. 246) This comes out, Flavell explains, in situations in which different people (or a single individual at different times) represent a single thing in several different ways – "ways that would be mutually contradictory if they described the object itself rather than mental representations

process in which the old theory (the situation theory) is completely replaced by the new and better theory (the representational theory), however: "The representational view does not supplant the situation theory, but only amends it for certain problems. Even as adults we remain situation theorists whenever possible and treat mental states as straight propositional attitudes." (1991, p. 252) This comes close to the current proposal of relational mindreading: as adults, our default interpretation strategy is to relate people to certain 'situations' that make up their reasons and their goals. In a recent paper, Perner and Roessler (2010) argue that young children explain other people's actions by appealing to evaluative facts about the external world. These children conceive of such normative facts as reasons that motivate an agent to adopt a goal and act correspondingly. Importantly, these motivating reasons are treated as fully 'objective reasons', "relativized neither to the agent's instrumental beliefs nor to her pro-attitudes." (p. 205) For the young child, that is, an agent's actions are explained by the facts that dictate how one ought to be motivated in the agent's situation; there is no room for personal deviation from this public norm. As a picture of adult goal-reason psychology, this characterization approximates the level of social understanding reached by our Rylean ancestors in the previous chapter. Yet, if we were to follow Perner's (1988, 1991) account, it would not involve any notion of mental representation.

Robert Gordon (1987, 2000a, 2000b, 2001) has also stressed the relational, or in his terminology, 'factive' nature of ordinary reason explanation. Gordon's characterization of reasons for action is broadly similar to my use of the term in the ordinary sense of goal-reason psychology: things in or about the world that favor adopting certain goals and performing certain actions. In Gordon's own words, a reason for action, in the strict sense, is "a favorable consideration, something about the world—a fact—that, at least to the agent's eyes at the relevant time, favored, or argued in favor of, doing what he did." (2001, p. 178) Giving an explanation of someone's action in terms of his reason that p, Gordon argues, normally goes accompanied by the presupposition that it is the case that p. Explanations of the form 'she a's because p' are 'factive' insofar that they commit the interpreter to it being the case that p. For a fact to be considered as the agent's reason, and thus for it to serve as an explanation of his action, "it must be a fact of which the agent is aware, a fact that is known to the agent…" (2000a, p. 77) Considering the fact that p as an agent's reason

of it." (ibid.) Later, approximately from 4 years onwards, children "gradually realize that people's cognitive connections engender inner, mental representations of their external objects, and that the same object can be represented in different, seemingly contradictory ways." (p. 247)

for action therefore not only presupposes that it is the case that p, but also that the fact that p *is a fact to* the agent.

The concept of knowledge at play here does not, however, presuppose the concept of belief. Attributions of knowledge in the relevant sense "bespeak not sophistication but rather *lack* of it – or, at least, failure to use the competence one has." (1987, p. 130-131, emphasis in original) This means that in default cases of reason attribution, interpreters simply confine the agent's knowledge to their own epistemic horizon, thereby failing to make allowance for the agent's possibly false or differing beliefs. Factive interpretation, Gordon argues, is our default strategy, "the form that is used unless one has some reason not to use it." (2000b, p. 105) Explicit belief explanations make sense when the factive implication is unwarranted. But this is exception rather than rule: "knowledge, attributed by default, is the normal epistemic condition of others; mere belief is the noted exception.' (Gordon, 1987, p. 132)

The state of knowledge attributed in default reason explanation is a relational state, a state that relates the agent to what the interpreter considers to be some reason-constituting fact. But on Gordon's account, relational states are not to be conceived as functional role states. In a paper directly targeting Sellars's Myth of Jones, Gordon invites us to consider another fictional ancestral tribe he calls the 'Outlookers' (2000b). Although the Outlookers never referred to mental states or episodes, "they gave appropriate explanations of action – causal explanations, it would appear, in terms of the reasons for which the actions were performed – though strictly in terms of public properties of public objects, or at least what they took to be public properties." (p. 105) Gordon calls them the Outlookers because "they were always looking outward to the world, never inward to the mind of the agent." (ibid.) For them, that is to say, "the mental is spread out over the world, coloring objects and situations with emotional and motivational charges, and not yet bottled into minds." (ibid.) The explanations the Outlookers gave of each other's actions were of the factive kind illustrated above, the kind Gordon argues we also tend to provide by default. The Outlookers were able to attribute default knowledge to one another in the course of reason attribution, but they lacked the sophistication we have "to speak *also*, when the need arises, of the mental causes of action." (p. 106, emphasis in original) Yet the similarity between their reason explanations and the ones we tend to give most of the time clearly suggests that "mental causes are in general a second best, invoked when there is reason not to locate the explanans out in the world." (p. 106)

The kind of knowledge Gordon claims we attribute to others in default cases of reason explanations is *not* to be conceived as a mental cause of the

action, then. The cause of the action is something out in the public world, not something 'bottled in the mind'. The knowledge attributed, it seems, merely connects the agent to the facts explaining his action; it is not itself part of the explanans.[60] Gordon's view on reason explanation is closely connected to his understanding of the Simulation Theory. According to Gordon, "it is simulation that makes it possible to think of people as acting from or because of reasons." (2001, p. 177) Simulating another person, he explains

> requires an egocentric shift, a recentering of my egocentric map on [the other]. He becomes in my imagination the referent of the first person pronoun "I," and the time and place of his [action] become the referents of "now" and "here." And I [...] cease to be the referent of the first person pronoun [...] Such recentering is the prelude to transforming myself in imagination into [the other] much as actors become the characters they play. (Gordon 1995, p. 55)[61]

Having transformed myself, in imagination, into the other, I then 'look out' into the world from the other's perspective and use my own response mechanisms and practical reasoning skills to identify the relevant reason-constituting facts. Thus far, the simulation routine does not make any reference to (FR-) representational states. And the subsequent attribution of the reasons identified to the other, it seems, does not require reference to such states either. The other can simply be interpreted as looking out onto the reason-constituting facts identified through simulation. Having the capacity to attribute default knowledge through simulation, Gordon argues, his Outlookers, unlike Sellars's Ryleans, were in no need of a Jones to teach them how to explain behavior in terms functional/causal role states.[62] The implication is that we do not need to

60 cf. the 'non-psychologistic' account of e.g. Dancy (2000), discussed in the appendix to chapter 2.

61 Gordon rejects the idea simulation is 1) an analogical inference from oneself to others, 2) premised on introspectively based ascriptions to oneself, 3) requiring prior possession of the concepts of the mental states ascribed (1995, p. 53). Contrast Goldman's version of the simulation theory in section 5.2 (see especially Goldman 2006, pp. 185-188). It should be noted that Gordon's rejection of the last point concerning concept possession primarily applies to simulation of emotions (cf. Gordon 2008) and proximal goal-directed actions (cf. Gordon 2005). It is does not seem to apply in the case of attributing propositionally articulated goals and reasons to others.

62 Gordon (2000b) presents his visit to the 'Outlookers' as an alternative to Sellars's Myth of Jones. But one wonders whether his Outlookers are really all that different from the Ryleans. Gordon's treatment of Sellars's myth seems to be premised on the idea that Sellars's Verbal Behaviorism (see section 3.1) attempts to *reduce* mind to behavior. On this construal of behaviorism, Gordon has a point when he states that Sellars's Ryleans "carry a much heavier burden than the restriction to public language. They are restricted to a much more austere idiom, which eschews not only causal explanations of human action in terms of mental states and episodes, but

refer to mental causes either when providing factive reason explanations of the actions of others, i.e. when engaging in relational mindreading.

The general idea of relational mindreading seems to allow for more than one interpretation. Both Perner's 'situational' attitudes and Gordon's 'factive' attitudes lack the (FR-)representational/causal dimension of Jones's teachings in Sellars's myth. In light of this, I propose a distinction between relational mindreading understood as consisting in (i) the attribution of *first-order* states and (ii) the attribution of *second-order* states. As we shall see, Perner's and Gordon's proposals are examples of the first option, whereas our discussion of Jonesian mindreading in the previous chapter suggested the second. Interestingly, however, Sellars's myth also allows for a first-order reading of relational mindreading.

4.3 Relational Ascent

Consider the *pre*-Jonesian Ryleans once more. John walks down the street and his Rylean friend asks him what he's up to. John replies: "I'm going to buy some milk at the supermarket." His friend asks him why. John answers: "The supermarket will be closed tomorrow." John has succeeded in giving his goal and his reason for his action (a) of walking down the street: his goal is to buy some milk at the supermarket and his reason is that the supermarket will be closed tomorrow. How does his Rylean friend interpret him? Let us focus on the attribution of John's reason and let his goal therefore be incorporated in the description of his action as "going to buy some milk at the supermarket."

also causal explanations of human actions *in terms of reasons* ..." (pp. 105-106; emphasis added) But as we have seen in chapter 3.1, the verbal behaviorist treatment of the Ryleans was only meant to reveal that their conception of mind was confined to public displays of intentionality. In fact, Sellars (e.g. 1953/2007, 1954/2007, 1969/2007, 1974/2007) tries to account for the meaningfulness of public language in terms of the inferential roles of linguistic expressions in 'the game of giving and asking for reasons'. Gordon moreover thinks that Sellars's story cannot account for the fact that ascription of propositional attitudes is systematically coordinated with their verbal expression, since 'outlooking' expressions only requires training in a public language, whereas 'inward-looking' ascriptions are guided by Jones's functionalist theory of mind. In general, he thinks that Theory Theory accounts cannot explain so-called 'Moorean paradoxes' such as S: "It is raining, but I don't believe it is" – a formally consistent sentence, assertion of which is self-defeating. TT-accounts, Gordon argues, fail to capture the paradoxical nature of sentences such as S: utterance of the assertion "It is raining" may be outweighed by other behavioral evidence that serves as input for the theory, so as to yield the ascription "I don't believe it is raining." Although this may indeed be true of present-day versions of TT, it is a mischaracterization of Sellars's myth. Chapter 3 clearly revealed that Jones's functionalist characterizations of mental states *are derived from*, and are thus *systematically coordinated with*, the Ryleans' pre-existent (first-order) understanding of each other's 'outlooking' linguistic utterances. See Rosenberg (2004/2007) for a detailed discussion of Gordon's (2000b) interpretation of Sellars.

It seems the interpreter has two options.

According to the first, she interprets him simply as follows: "John is going to buy some milk at the supermarket because the supermarket will be closed tomorrow," or "John is a-ing because *p*." On her verbal behaviorist understanding, this would mean "John is a-ing because *p-out-loud*(J)," where '*p-out-loud*(J)' means something like '*p* out of John's mouth'. Of course, for John's answer that *p* to make sense to her, it has to be framed in an inferentially articulated context, e.g. that one can buy milk at the supermarket, that John's brought some money to pay for the milk, that he will reach the supermarket before the end of the day, etc. For the Rylean interpreter, this means that John should have access to the world and the inferences that it licenses if his answer that *p* is to make proper sense of his action, access evidenced by e.g. his giving proper answers to subsequent questions. The mental state '*p-out-loud*' attributed to John is what I will call a *first-order* relational state: it is the state of *being-directed-at* what is specified by the content clause that *p*, in this case: that the supermarket will be closed tomorrow.

On the second option, John's friend interprets him not (only) as "John is a-ing because *p-out-loud*(J)" but (also) as "John is a-ing because he *says that p*" or "John is a-ing because he *thinks-out-loud that p*." For his friend to interpret him in this way, she needs to use their semantical meta-language, i.e. she needs to ascend from attending to *what* he's saying or 'thinking-out-loud' to attending to *his saying it* or '*his thinking-it-out-loud*'. As we have seen in chapter 3, engaging in semantical discourse in this manner allows the Ryleans to comment upon or criticize each other's sayings, and to do so with a generally functionalist conception of such sayings in mind: how, in accordance with further specifiable entry/inference/exit rules, a certain utterance type is supposed to function in their linguistic practice. A shift towards Rylean semantical discourse is a form of what Quine (1960) called 'semantic ascent': "the shift from talking in certain terms to talking about them." (p. 271) Since this functional classification of their language use in semantical discourse amounts to the FR-concept of mental representation, I shall refer to this shift to Rylean semantical discourse as an instance of 'functional role representational ascent' or FRR-ascent.[63] On this second option, then, John's friend makes an *FRR-ascent* from "John is a-ing because *p-out-loud*(J)" to "John is a-ing because *he thinks-out-loud that p*." The mental state 'thinking-out-loud that *p*' is what I shall term

63 Using Sellars's dot-notation (see appendix chapter 3), the difference between the first and the second option is the difference between "John a's because *p-out-loud*(J)" and "John a's because an overt ·p· (J)," or between "John a's because *p*(J) and "John a's because a ·p· (J)."

a *second-order* state: it is the state of the agent of *being-in-a-state-about* what is specified by the content clause that *p*.

Let us now apply the same considerations to John's friend's *post*-Jonesian interpretation of his action. Strictly following Jones's teachings, John's friend would interpret the reason he issued in his response "because the supermarket will be closed tomorrow" as being caused by an inner episode of thinking-to-himself that the stores will be closed tomorrow. On this post-Jonesian silent counterpart of the second pre-Jonesian option discussed above, she would interpret him as "John a's because *he thinks-to-himself that p*," where 'thinks-to himself that *p*' is a *second-order* FR-representational (FRR-)state. Following the first pre-Jonesian option, however, she would interpret John as "John a's because *p-to-himself*," where the state attributed is the 'offline' counterpart of the *first-order* overt state '*p-out-loud*(J)'. Applying FRR-ascent to this first-order state of '*p-to-himself*', we get the second-order FRR-state of '*thinking-to-himself that p*'.

The discussion of Sellars's myth in the previous chapter pointed in the direction of the second option for Jonesian mindreading. And perhaps this is also the best way to read Sellars's myth. For if John's friend is indeed supposed to interpret John's utterance that *p* strictly according to the teachings of Jones's theory, i.e. *as being caused by* its silent counterpart, it seems she would have to make an FRR-ascent: the semantically characterized FRR-state of thinking that *p* according to specifiable entry/inference/exit rules could then also be regarded as the theoretical causal role state specifiable in causal patterns mirroring these rules (see chapter 3.4). Without FRR-ascent, it is not clear whether the interpreter would have enough conceptual material to consider the attributed state as a causal role state: there would be no functional classification to exploit as a model for the causal role to be assigned to the relevant state. But even though the second option may be closer to Jones's teachings in Sellars's myth, there appears to be no reason why the first option could not have been available for Jonesian mindreaders as well.

Notice first that even in their post-Jonesian stage, the Ryleans would still seem to be able to make perfect sense of each other's *overt* utterances in their *pre*-Jonesian style. There seems to be no gain for his Rylean friend to interpret John's utterance "because the stores will be closed tomorrow" as being caused by an inner episode with the functional role of saying that the stores will be closed tomorrow. For in order to give a useful characterization of this functional role, she would first have to understand the utterance classified by it, the very utterance John utters in response to her question.[64] And this is

64 Recall that on Sellars's story, the Rylean's interpretation of each other's utterances as lin-

something she could already do prior to Jones's teachings. So why not simply skip the reference to an inner episode and the FRR-ascent required to specify its causal role?

But now consider interpretation of behavior without the guidance of the agent's overt utterances. According to the second option, the interpreter is supposed to reason as follows: "this behavior of the agent would make most sense if he thought-out-loud that *q*, so the inner episode causing the behavior is likely to be an inner 'thinking-to-himself that *q*'." But why couldn't she simply think: "this behavior of the agent would best make sense in the presence of '*q-out-loud*(A)' so A *q's-to-himself*?" That is: why apply FRR-ascent and go through the trouble of characterizing the causal role of the posited inner episode, if she can model the inner episode directly on her *first-order* understanding of the overt utterance that would best fit her behavior? Presumably, this would not be the way in which the Ryleans initially learned to engage in Jonesian mindreading. But couldn't they train themselves, through practice, to make a shortcut on Jones's theory?[65]

To give an example: suppose John's Rylean friend knows his daily routine and hence knows that John goes to the supermarket every day to buy some milk because he ran out of the milk he bought the day before. Instead of strictly following Jones's theory and model the causal role state posited to explain his walking down the street on the FR-representation of the overt utterance that would have made most sense of John's behavior, she would interpret John's walking down the street as "a-ing because (I ran out of milk)-*to-himself*," thus using her Jonesian mindreading techniques without FRR-ascent and, it would seem, also without considering his mental state as having a causal role in the explanation of his behavior.

On this reading of Jonesian mindreading, interpreting John as going to the supermarket because he ran out of milk would consist in 'hearing' or having 'heard' him say to himself that he ran out of milk, analogous to the way

guistic acts does not rest on their ability to engage in semantical discourse (see chapter 3.5 and appendix to chapter 3). Rather, FRR-ascent appears a useful way for Rylean interpreters to classify each other's linguistic performances for further purposes, e.g. to criticize such performances or, indeed, to provide causal explanations of each other's behavior. See also chapter 5.2.

65 It should be noted that attribution of *both* the first-order state "because q-to-himself" and the second-order state "because he though-to-himself that *q*" could either (a) be the result of drawing a *theoretical inference* from observable behavioral phenomena to the occurrence of the thought episode, or (b) be an instance of making a non-inferential, *perceptual report* about the occurrence of the episode. As explained in chapter 3.5, Sellars's treatment of mindreading allows for both (a) and (b). In the case of first-order state attribution, (a) would be an inference from certain behavior to '*q-to-himself*' and (b) would be the perception of '*q-to-himself*' *in* behavior. Correspondingly, second-order state attribution would involve (a) an inference from behavior to 'thinking-to-himself that *q*', or (b) perceiving 'thinking-to-himself that *q*' in the behavior manifested.

pre-Jonesian interpretation could proceed by hearing or having heard him say "I ran out of milk," i.e. without engaging in semantical discourse in order to classify his utterance in functional terms. It would be analogous only to the extent of being semantically evaluable, of course, not necessarily in being accompanied by inner sounds or 'the wagging of inner tongues' (cf. chapter 3.2). Just as the Ryleans would have been able to learn to communicate *in* a second language after first having gone through a phase of thinking *about* it (by matching, in their meta-language, the functional roles of utterance types of the foreign language to the functional roles of utterance types of their own language in order to figure out what the foreign utterances mean), they could also have learned to think about other people's mental lives directly in a 'mental language' rather than always having to think *about* it in the FR-representational terms of corresponding overt linguistic acts (see appendix chapter 3).[66]

At this point, I do not wish to advocate either of the two options identified above, i.e. (i) relational mindreading by merely attributing first-order states or (ii) relational mindreading by (also) attributing second-order states, such as FRR-states through FRR-ascent. Prima facie, both options seem to be available as a further specification of the general idea of relational mindreading. And even on Sellars's fictional story, it is perfectly possible that the Ryleans used both techniques in daily social interaction. Rylean interpreters would have been prone to making FRR-ascent especially in those situations in which other people's overt linguistic acts (or the occurrence of their silent counterparts) did not seem to accord with proper language use. The use of FRR-ascent would have enabled them to criticize each other's apparently improper thoughts and utterances. In such relatively problematic social situations, commenting upon each other's thinkings and sayings in terms of second-order FRR-states would also have invited them to think about the mental causes of each other's behavior and to seek further explanations as to why someone acted in the strange manner he or she did. In more spontaneous and less problematic social situations, interpretation could then have proceeded according to the first option, by merely attributing first-order states to one another.[67]

66 Of course, this 'mental language' would not be a private language, in Wittgenstein's (1953) sense, for it would be modeled entirely on the Rylean's public language.

67 A causal analysis of reason explanation is often combined with a BD-Model of action explanation. The first thing to notice is that the causal analysis of reason explanation is neutral with regard to the question whether the states attributed are genuine representational states or merely *second-order* relational states. Jones taught the Ryleans to attribute second-order FRR-states to one another: relational, yet causally relevant 'inner' states (see chapter 3.4). Perner and Roessler (2010, pp. 206-210), however, argue that a causal analysis of reason explanation need not refer to

Let it also be noted in passing that the distinction between first-order and second-order mental states is not wedded to a particular Sellarsian conceptual framework. As we shall see in the next section, one could replace Sellars's FR-conception of representation with some other representational notion and still maintain the distinction between a relational conception of another agent as (i) being directed at the world and (ii) being directed at the world in virtue of internal representations of the world. In general, the distinction hinges on the question whether or not an agent's intentional attitudes towards and her interaction with the environment should be conceptualized as being mediated by mental representations. A first-order conception of mental states implies a negative answer to this question, a second-order conception an affirmative one. Sellars's myth allows us to make this distinction at the level of *folk psychology*. But it also seems to apply at the level of *cognitive science*. The current debate about the status of mental representations in (the philosophy of) cognitive science comes down to the question whether it pays off explanatorily to conceive of the internal neural processes underlying interaction with the environment as representations of the worldly offerings the organism that instantiates these processes is directed at (cf. Ramsey 2007). Proponents of the embodied cognition and enactivist paradigms lean towards a negative answer to this question (e.g. Menary 2006, Gallagher 2008a, Hutto 2008a, Hutto 2011b, Thompson 2007, Chemero 2010), whereas defenders of the more classical cognitivist paradigm stick to a positive answer (e.g. Fodor 2008, Bechtel 2008).

With the distinction between first-order and second-order relational mindreading in place, we can read Ratcliffe's objection to the idea that reason explanation in terms of situations additionally requires reference to mental states as an objection specifically directed at interpretation conceived in terms of *second-order* relational attitudes, e.g. FRR-states causing the action to be explained. Perner's proposal can be interpreted as saying that young children (and adults, by default) interpret other agents in terms of *first-order*, *non*-representational propositional attitudes, states relating them to the situations that constitute what he calls their 'objective reasons'. Gordon's Outlookers can now also be regarded as first-order relational mindreaders. The relational state

inner *mental* causes, but could also be given in terms of the outer, worldly causes of action. They suggest that causal explanations advert to facts that 'make a difference', where this is spelled out in terms of counterfactual dependence along the lines of J. Woodward's (2003) 'interventionist' analysis of causal relations. Crudely, the idea is that an 'objective reason', i.e. some worldly fact, causally explains an action in the sense that, were there to be an intervention on the facts that give someone a practical reason, there would be a corresponding change in her action. It seems this analysis of reason explanation is also available on a *first-order* understanding of relational mental states.

of knowledge attributed to another agent when explaining her action in terms of the factive reason that p can be modeled as the agent's silent assertion that *p*, i.e. as '*p-to-herself*'. Attributing such first-order state, an interpreter would merely conceive of the agent as being intentionally directed at the explanans of the action out in the public world, i.e. the fact that *p*; he would not also understand the action as being caused by a representational state about *p*.[68]

4.4 Metarepresentation Is Not Enough

It is time we ask ourselves a question that has been staring us in the face ever since Jonesian mindreading was revealed as being restricted to the relational level of discursive understanding, somewhere along the way in chapter 3. By carefully following through Sellars's conceptual bootstrapping strategy in his Myth of Jones, we discovered that the move from a verbal behaviorist towards a functionalist conception of mind doesn't add up to a genuine understanding of belief and desire, an understanding on which one person's take on the world may defy another's. Jones, that is, didn't bootstrap the Ryleans into representational mindreading. The question, then, is: How could these interpretative shortcomings of the educated Ryleans be alleviated? What should be added to their social skills in order to transform them into mature representational mindreaders? What does it take, for example, to make them understand differing beliefs, beliefs that might turn out false?

One important lesson from chapter 3 was that mental states can be conceived as inner representational states without thereby being understood as subjective or private representational states, i.e. states that specify the way the world appears to a specific individual, and not necessarily to anyone else. It is the latter conception that is required for genuine understanding of false beliefs, commitments as to how things are in the world that are in conflict with commitments of, and hence not shared by, the interpreter herself. Jones taught the Ryleans how to apply their preexistent understanding of each other's overt utterances as FRR-states in order to characterize the posits of his new theory of mind. But by doing so, he didn't take away what we might call a *disjunctive constraint* on their interpretation skills.

Recall that even in their post-Jonsesian stage, the Ryleans could only interpret the thinking(-out-oud) of a false thought that *p either* as a failed attempt to

68 Relational ascent should not be conflated with Gordon's *ascent routines* for propositional attitude ascription (1995, 1996, 2000b, 2007). See appendix for discussion.

perform the world-directed mental act of thinking that *not-p*, *or* as a successful attempt at indicating a mere possible scenario according to which *p*. They could not grasp the idea of a speaker or thinker considering such counterfactual scenario *as actual*; they could not appreciate the fact that other people's attitudes towards the world could be informed by false representations of the world. The Ryleans were disjunctivists of sorts. Their conception of mind did not allow for an understanding of both true *and* false claims/thoughts as being sincere, world-directed propositional attitudes. Such 'common factor' understanding of true and false (overt) thoughts would require representational mindreading.[69] For the educated Ryleans, world-directed propositional attitudes incompatible with the public take on the world were still unintelligible. Jones's theory didn't sever the bonds of public assessment that constrained pre-Jonesian mindreading. In their educated state, the Ryleans were still oblivious to this private dimension of the mental (chapter 3.3).

It may appear as if this shortcoming of Jones's theory is a consequence of the particular notion of representation he exploited, the FR-notion. Recall that on this FR-understanding, a mental representation is merely some symbolic vehicle that ought to indicate something in the world (or some possible scenario), according to the entry/inference/exit rules that define its use in linguistic practice. Importantly, the indicating vehicle need not show any resemblance to what it is supposed to indicate. But what if Jones had used the idea of a *model* as a model for his theory? In contrast to symbolic representations, models (maps, pictures, miniatures, etc.) represent in virtue of sharing certain structural features with their representational targets. With respect to those features, a model is supposed to *match* or *correspond* with its target.[70] Suppose that Jones had added to his theory something along the following lines: "Overt utterances about the world or their silent counterparts are actu-

69 One could draw a parallel here with disjunctivist theories of perception (e.g. Byrne and Logue 2009). The difference is that disjunctivist theories of perception are typically concerned with the nature of perceptual experience, whereas our (Sellars's) Jonesian account would (merely) target the folk *concept* of perceptual experience. Accordingly, our basic (i.e. relational) commonsense understanding of perceptual experience would be of a disjunctivist kind. It would only be from a more reflective stance that we stop being 'naïve realists' and start conceiving of our perceptual contact with the world as being mediated by our subjective and possibly misleading impressions.

70 Like the functional role notion of mental representation, this model notion is also widely used in the philosophy of mind and (the philosophy of) cognitive science. See e.g. Braddon-Mitchell and Jackson 1996, O'Brien and Opie (2004, 2011), Ramsey (2007). Millikan (e.g. 2004) uses the model notion within her teleosemantic account of mental content. Sellars also used a 'picture' notion of representation, but not to characterize the folk concept of mind. It plays an important role in his attempt to fuse the 'manifest' or commonsense image with the 'scientific' image of man (Sellars, 1963/1991). For discussion of this aspect of Sellars's philosophy see e.g. Millikan (2005, ch. 4), deVries (2005, ch. 2), O'Shea (2007, ch. 6), Rosenberg (2007, ch. 5).

ally about *an internal world that is about* the world, a *model of* the world *projected onto* the world like a movie shown on a screen; it is the resemblance between the internal model and the external world that ensures successful exploration of the world through thought, talk and action." With this addition to Jones's theory, it may seem as if the Ryleans would have been given the necessary tools for representational mindreading. By interpreting each other's behavior as being caused by mental models of the world, the possibility of a mismatch between model and world may seem to come within the purview of their social understanding.

But this is a mistake. The mere positing of internal miniature worlds projected outwards, does not enable Jones to remove the disjunctive constraint on Rylean interpretation as outlined above. Let us take a step back and ask ourselves again why Jones's original theory didn't take the Ryleans beyond a relational conception of one another. As we saw in chapter 3.3, the reason was that application of his theory of thought relied entirely on the Rylean's pre-existent verbal behaviorist understanding of each other's overt linguistic acts. The functional roles of inner episodes were derived from pre-Jonesian meta-linguistic specifications of Rylean language use, specifications made exclusively with reference to what was publically accessible, out in the open for everyone to see. The same considerations apply to the modification of Jones's theory currently under discussion. Jones's teachings still depend entirely on the Rylean's pre-existent linguistic understanding. As before, inner episodes are modeled on overt linguistic acts. The only difference is that he now also introduces an inner world, modeled on the outer world. Again, however, *merely moving the game inside does not sever the bonds of public assessment*. For the Ryleans, the internal models posited by Jones as representational intermediaries are simply copies of the world outside, which still serves as the ultimate and only reference when it comes to interpreting other people's world-directed attitudes. On their new understanding of mind, utterance of the false sentence that *not-p* would be interpreted *either* as being about a *malformed* model of it being the case that *p*, *or* as being about a *replica of a counterfactual scenario* in which *not-p*. Addition of a 'mental model' conception of representation by itself does not extend to the idea that other people's attitudes towards the *actual* world *can be informed by* representations of *counterfactual* scenarios.

What we need is a shift from understanding the speaker as representing a counterfactual scenario towards understanding her as representing the actual world in counterfactual terms. In the case of the Ryleans, this requires that a distinction be made between how one *ought* to be representing the world and how one actually *is* representing the world. According to the rules of

their linguistic practice, utterance of a false sentence ought to be directed at a counterfactual scenario. What the Ryleans still have to learn is that the way an individual person conceives of the world may defy these public, intersubjective oughts, that the utterance of a false sentence may give expression to the speaker's subjective view on the world. This *subjective* dimension of mind cannot be accounted for simply by conceiving of mental representations as 'inner'. The public-private dimension of mind runs orthogonal to the distinction between 'inner' mental states and their 'outer' manifestations.

Most participants in the debate on folk psychology remain silent on what concept of mental representation they think is involved in representational mindreading. Perner (e.g. 1988, 1991, 1995) has been a notable exception. On Perner's earlier account (1988, 1991), mental representations are modeled as mental models, as suggested above. Acquisition of a representational theory of mind around the age of 4 requires that children come to represent other people's mental models as such, i.e. that they become meta-representers in Pylyshyn's (1978) sense of the term, representing the representation relation between the agent's mental model and what it models. As we have seen in section 4.2, Perner holds that children start out as 'situation theorists', 'looking through' the agent's representational medium and directly relating the agent to the situation attended to. Around 4 years of age, however, they come to understand that the agent's intentional directedness is actually mediated by internal models *of* what it is directed at. This requires that they distinguish between what is being represented – the situation attended to – and how the agent represents it – the content of the model of the situation. And this, Perner argues, should allow them to understand that mental states may guide action in the real situation (what is being represented) as if it were a different situation (how the model represents it), i.e. to understand false belief.

In later work (e.g. 1995), Perner proposes an analysis of mental representation along Fodorian lines, according to which to believe that p is to stand in a relation of semantic evaluation to the proposition expressed by a token mental representation.[71] Situation theorists can semantically evaluate propositions, but cannot conceive of the fact that propositions are semantically evaluable. Thus they simply relate others to what they evaluate as true (real situation) or false (pretend situation). What children learn around at 4 years

71 Cf. Fodor's (e.g. 1987) formulation of the Representational Theory of Mind (RTM): "For any organism O, and any attitude A toward the proposition P, there is a ('computational'/'functional') relation R and a mental representation MP such that: MP means that P, and O has A iff O bears R to MP (p. 17). Perner's modification is to conceive of R not in terms of computation but rather in terms of semantic evaluation.

of age, is to regard others as semantically evaluating the proposition expressed by a mental representation, i.e. as evaluating the proposition expressed as true or false. Children can now distinguish between what is being said about a situation – the proposition expressed – and the situation itself – 'of which' the proposition is evaluated as true or false. And this, again, should allow them to understand that others may assign a different truth value to a proposition expressed than the one it has, as in the case of false belief.

In many interesting ways, Perner's developmental account runs parallel to our discussion of Sellars's Myth of Jones. Perner's situation theorists suffer from a similar disjunctive constraint on their social understanding as the Ryleans. They can entertain multiple models, mark some as merely as-if and relate others to the (as-if) situations modeled, but they cannot understand that others may have a mismatching model of the real situation at hand. The situation theorists' concepts of belief and pretence are confined to respectively 'acting on a true proposition' and 'acting on a false proposition'; they cannot understand that others may act on a false proposition they evaluate as true. These primitive concepts, Perner argues (1995, see also Perner, et al. 1994), can be regarded as two sides of one single concept of 'prelief'. Understanding others in terms of 'prelief', they either adopt an attitude of 'holding true' towards true propositions or adopt an attitude of pretence towards false propositions. This easily extends to the disjunctive constraint as explicated above: if someone's action is informed by a false proposition, she must either be engaged in pretend play or have failed to act in a 'holding true' frame of mind.

I merely take issue here with Perner's suggestion that the capacity to metarepresent, i.e. represent the mental representation relation, *suffices* to ascribe false belief. Understanding of someone's intentional directedness towards the world as merely being mediated by a model of the world does not amount to an understanding of that model as a *subjective* model of the world, i.e. as not necessarily shared by interpreter and interpretee. The ability to conceive of someone as semantically evaluating a proposition does not by itself endow one with the capacity to interpret that person as evaluating the proposition *other than she ought to*.

On Sellars's 'deflationary' analysis of truth in terms of 'semantic assertibility', for example (e.g. 1967/1992, ch. IV)[72], the function of truth statements of the form 'that *p* is true' is, roughly, to explicitly authorize the performance of asserting that it is the case that *p*, i.e. to comment upon the move towards

72 This analysis of truth in terms of semantic assertibility should not be conflated with Sellars's complementary 'picture' theory of truth (e.g. 1967/1992, ch. V – see footnote 70).

saying(-to-oneself) that it is the case that p as being in accordance with the entry/inference/exit rules that define ·it is the case that p·s in the home language (cf. appendix chapter 3). Accordingly, semantically evaluating a proposition only requires ascent to semantical meta-language (cf. chapter 3.2 and chapter 4.3). In theory, it is something that Perner could teach his situation theorist without teaching them our mature concept of belief. On this proposal, saying that a proposition is true (false) is merely a means of making explicit that asserting that proposition is (im)permissible according to the 'public oughts' of linguistic practice. Similar results are obtainable using other deflationary analyses that explain truth exclusively in terms of the expressive role of truth-talk in discursive practice (e.g. Brandom 1994, 2002). There may be other concepts of truth, proper use of which does require the capacity for representational mindreading (see chapter 6.4). The point here is merely that acquisition of the concept of semantic evaluation can be cashed out in relational terms. Without further specification, Perner cannot use it to characterize the ontogeny of representational mindreading.[73]

So it seems that the conclusion from the chapter 3 generalizes: mere metarepresentational interpretation, whether in terms of FR-representations, mental models or semantically evaluated propositions, does not suffice for an understanding of others as subjectively representing the world to themselves, and hence does not suffice to remove the disjunctive constraint on Rylean interpretation.

Let it also be noted that the distinction between relational and representational mindreading cannot be framed exhaustively in terms of the linguistic contrast between extensional/transparent and intensional/opaque contexts. The sentence S 'There is a bottle of water in the fridge', has an extensional or transparent context, such that the truth of S entails that there is a bottle of water in the fridge and that the truth-value of S is not altered by substitution of co-referential terms (e.g. 'water' by 'H_2O). Accordingly, on an extensional interpretation of John's utterance (U) of S, U would imply that there is a bottle of water in the fridge and substituting 'water' for 'H_20' would not render the interpretation inadequate. The sentence S* 'John believes that there is a bottle of water in the fridge', by contrast, has an intensional or opaque context. S* does not entail that there is water in the fridge (John's belief may be false) and substitition of co-referential terms 'water' and 'H_20' is not automatically truth-

73 Perner is of course well aware of the fact that belief ascription requires sensitivity to this subjective dimension of propositional attitudes. My point is merely that the subjectivity of belief cannot be analyzed in terms of some or other notion of representation, as Perner seems to suggest.

preserving (John may not know that water is H_20, so that the sentence 'John believes that there is a bottle of H_20 in the fridge' may be false). Carried over to our understanding of John's utterance U 'There is a bottle of water in the fridge', an intensional interpretation of U does not imply that there is in fact a bottle of water in the fridge (U may be based on a false belief) and substitution of co-referential terms 'water' and 'H_20' may render the interpretation inadequate (for John may not know that water is H_20). The intensional interpretation captures the way in which John's propositional attitude expressed by U represents the referent under a certain description, a description that may turn out to be false.

In order to clarify which elements of the content clause of an expressed propositional attitude are supposed to be understood intensionally and which elements extensionally, one could divide the content clause into two parts, an extensional 'de re' part marked by 'of' or 'about' and an intensional 'de dicto' part marked by 'that'. Thus, depending on the situation and what one knows about John, one may infer from U that, e.g., John believes *of* (about) the bottle in the fridge that there is still water in it (one knows, for example, that the bottle has been refilled with milk since John last saw it). The extensional, *de re* part of the ascription is removed from the content clause of the ascribed belief and reflects the ascriber's own attitude towards what it refers to; the intensional, *de dicto* part of the ascription captures the way in which the extensionally specified referent is interpreted as being conceived by the ascribee.[74]

Now there is no reason to assume that Jonesian mindreading only allows for a *purely extensional* understanding of other people's minds. Suppose six-year-old John thinks that there is a bottle of water in the fridge. His Rylean caregivers know perfectly well that they should not ascribe to John the thought that there is a bottle of H_20 in the fridge. This is not something six-year-old John is disposed to meaningfully think-out-loud (or, derivatively, think-to-himself) under any circumstance. Similarly, the Ryleans would have no trouble recognizing *experts* amongst them. Such experts would give descriptions of worldly things that go beyond their own knowledge; learning something

74 Cf. Brandom's (1994) treatment of the de re/de dicto distinction: "The suggestion is that the expressive function of *de re* ascriptions is to make explicit which aspects of what is said express substitutional commitments that are being *attributed* and which substitutional commitments are being *undertaken*. The part of the content specification that appears within the *de dicto* 'that' clause is limited to what, according to the ascriber, the one to whom the commitment is ascribed would (or in a strong sense should) *acknowledge* as an expression of what that individual is committed to. The part of the content specification that appears within the scope of the *de re* 'of' includes what, according to the *ascriber* of the commitment, but not necessarily according to the one to whom it is ascribed, is acknowledged as an expression of what the target of the ascription is committed to." (p. 506)

from the experts would amount to acknowledging that what they have to say about these things reveals part of their hidden nature. In general, the Ryleans can conceive of people's knowing and not-knowing (remembering or forgetting, slower or faster reasoning, etc.) as cases of better or limited (restored or degraded, slow or upgraded, etc.) access to the public world and the inferences that it licenses. The phrase 'access to' is extensional, yet the adverbial qualifiers make it partly intensional. Jonesian mindreading, that is to say, allows for understanding *different points of view* on a publically accessible world.

But at the same time, Jonesian mindreading cannot accommodate a *wholly intensional* conception of propositional attitudes either. The different 'modes of presentation' in terms of which interpreters make sense of each other *must at least be compatible* with their own ways of conceiving things. The disjunctive constraint on their interpretation skills reveals itself as soon as a speaker makes a claim *incompatible* with public assessment. When this happens, the speaker can only be interpreted as either having failed to make a claim, or as indicating something about a mere counterfactual or imaginary scenario. The authoritative or persuasive speaker could of course try to *indoctrinate* her Rylean audience in order to cause them to change their views. But neither the speaker nor the audience could interpret the event as a case of *discarding* belief in light of new considerations. Jonesian mindreading does not allow for an understanding of *incompatible views* on the world 'from the same point', so to say. Genuine ascription of false beliefs is not an option.

Jonesian mindreaders are blind to the subjective or private dimension of each other's propositional attitudes. This deficit cannot be captured solely in terms of some or other concept of mental representation that the Ryleans would lack. Add any notion of representation to their mentalistic repertoire and their social understanding would still be confined to public evaluation. Nor can it be framed as a purely extensional understanding of each other's propositionally articulated directedness towards the world. For in certain important respects, they could exhibit a nuanced appreciation of the intensional aspects of each other's propositional attitudes, in terms of degrees of epistemic access to the inferentially articulated public world. These traditional contrasts between representation and metarepresentation, extensionality and intensionality, *de re* and *de dicto*, all fail to capture the essential feature of the Ryleans' social impairment: the inability to acknowledge that there is a private dimension to the public life of minds.[75]

75 Brandom (1994, ch. 8) claims that the representational dimension of discursive practice can be explained by offering an account of the use of the *de re*/*de dicto* distinction in the attribution

Mastery of our mature concepts of belief and desire consists in the capacity to attribute *subjective* representational states. The relevant contrast is between relational mindreading and *subjective* representational mindreading. Psychologically, the difference lies in the *information available* for interpretation of someone's world-directed attitudes. The disjunctive constraint on relational mindreading has it that a speaker's utterance that *p* when in fact (according to the interpreter) *not-p* can *either* be interpreted as a failed attempt at claiming that *not-p*, *or* as a successful attempt at making the supposition that *p*. Relational mindreading does not enable the interpreter to regard the *failed* claim that *not-p* as a *false* claim that *p*, expressing the speaker's false belief. In order to understand the utterance of a false sentence as the making of an empirical claim, one needs to take *what* is represented by the false sentence, i.e. some counterfactual scenario, *as pertaining to the actual world*. But this is impossible for the mere relational mindreader: not only is there nothing she can find in the actual world to relate the speaker to, but all that she could find there plainly contradicts what the speaker is saying. Contents incompatible with her take on things cannot be considered as the contents of world-directed attitudes, only as the contents of suppositional attitudes.

Representational mindreading takes away this informational constraint: it enables the interpreter to incorporate information incompatible with public evaluation into the content clauses of other people's world-directed attitudes. In our example, the representational mindreader can take the speaker's successful attempt at indicating the counterfactual scenario that *p as* a failed attempt at claiming that *not-p*, thus turning it into the *false* claim that *p*. On this understanding, the speaker can succeed in performing a *world-directed* linguistic act *by* talking about the counterfactual. What is represented by the speaker's

of commitments to others (cf. note 16). If my observation is correct, however, the mere use of this distinction does not evince a subjective understanding of other minds. According to Brandom, "…to grasp the representational content of [the] claims [of others] … is just mastering the social dimension of their inferential articulation – the way in which commitments undertaken against one doxastic background of further commitments available for use as auxiliary hypotheses can be taken up and made available as premises against a different doxastic background." (p. 517) What I have been trying to show, is that the (educated) Ryleans, though oblivious to the subjective character of mental representation, *did* master this social dimension of the inferential articulation of the contents of each other's (overt) mental states, but *only* insofar as the 'doxastic backgrounds of further commitments' against which to assess these states could be considered *compatible* with their own. Brandom's analysis of representational locutions does not by itself account for the *subjective* dimension of the practice of giving and asking for reasons. Brandom claims that "Beliefs and claims that are *propositionally* contentful are necessarily *representationally* contentful because their inferential articulation essentially involves a social dimension." (ibid., p. 519, emphasis in original) On a subjective understanding of the notion of representation, I contend, this claim is false. The intersubjective dimension of discursive practice does not necessitate a subjective understanding of other minds on the part of the participants of that practice.

utterance, i.e. the counterfactual scenario that *p*, can now be considered as a successful expression of her personal take on what she *ought* to have said, i.e. that *not-p*. Criteria for private success in taking a stand towards the world become differentiated from the dictates of public oughts.

(A short not on terminology: Up to this point, I have been using the term 'representational mindreading' to indicate *non*-relational interpretation of other people's goals and reasons in terms of genuine (false) beliefs and (discrepant) desires. In light of our present discussion, this may start to sound a little confusing. What I mean by the term, and have meant by it all along, is what I should now perhaps refer to as 'subjective' or 'S-representational mindreading'. This explicitly distinguishes interpretation in terms of our mature, subjective concepts of belief and desire from second-order relational mindreading using some or other concept of mental representation. From here on, I shall therefore add this qualifier.)

In light of the previous section, we can distinguish between two ways in which the 'subjective shift' from relational mindreading to S-representational mindreading can be accomplished: by adding a subjective dimension to (i) first-order relational states and (ii) second-order relational states. Suppose it is not raining but John believes it is. His (silent) utterance 'It is raining' can be regarded as an FR-representation, the function of which is to indicate conditions of rain (or some other kind of representation with the content 'It is raining'). What is represented by John's utterance 'It is raining' is the proposition that it is raining. But in the present circumstances, John ought to have said and thought that it is not raining. His representation of the world as the possible scenario in which it is raining misrepresents the fact that it is not raining. On this reading, then, what John's representational state 'thinking that *p*' is about, a scenario in which *p*, is taken as John's mistaken subjective view on what his thought ought to have been about: that *not-p*. The alternative is to replace the second-order state of 'thinking that *p*' by its first-order counterpart '*p-to-himself*'. On this first-order reading of S-representational mindreading, what John's '*p-to-himself*' is directed at, a *p*-scenario, is regarded as his misrepresentation of what he ought to have been directed at: the fact that *not-p*. On both options, the shift to S-representational mindreading enables the interpreter to understand the agent's assessment of the actual world in counterfactual terms, thereby adding a subjective dimension to the interpreter's social understanding. In general, the things the agent tries to accomplish (i.e. her goals) and the things that make them accomplishable and worth accomplishing (i.e. her reasons) can now be assessed from the agent's personal point of view, as the things she intends to do out of her desires in light of the things she believes.

The two options for S-representational mindreading are presented in figure 1.

	Relational mindreading	**S-representational mindreading**
First-order	State attributed to A: 'being-directed-at' goal/reason	*Subjective shift:* 'being-directed-at' goal/reason *as conceived by A*
Second-order	*Relational ascent:* 'being-in-a-state-about'… (e.g. FRR-state) goal/reason	*Relational ascent/subjective shift:* 'being-in-a-state-about'… (e.g. FRR-state) goal/reason *as conceived by A*

Figure 1: relational ascent and subjective shift

In chapter 3.3 I hinted at an ambiguity in our use of the notion of mental representation, an ambiguity that stands in the way of genuine appreciation of the social predicaments of the post-Jonesian Ryleans. I explained that the Ryleans already had a concept of mental representation before Jones came along, a concept that Jones relied on in teaching them his new theory, i.e. the FR-notion. Yet this notion of mental representation did not enable them to go beyond the public confines of relational mindreading. It did not enable them to enter the private dimension of other minds through S-representational mindreading. What we have seen in this section, is that this ambiguity in the notion of mental representation does not so much lie in the fact that there are two kinds of mental representation that, when conflated, render the term ambiguous. Rather, it has to do with the fact that we often *add* a subjective element to whatever kind of mental representation we are talking about, and that we do so without realizing it. Now that we have made this subjective dimension explicit, it should become easier to at least appreciate the *possibility* that there lies fault in the habit of projecting our sophisticated understanding of folk psychological practice in terms of beliefs and desires onto ordinary goal-reason psychology (see chapter 2.5). For now that we see that the concept of mental representation *simpliciter* does not *entail* the subjective character of our mature concepts of belief and desire, it should become apparent that it is

at least *conceptually* possible that the two might come apart, and hence, that framing commonsense goal-reason psychology exclusively in terms of belief-desire psychology may indeed be a fallacy.

4.5 Conclusion

In chapter 1 I identified two challenges for the Relational Model of folk psychology: to show that the distinction between attributing relational propositional attitudes and attributing S-representational propositional attitudes, i.e. the distinction between relational mindreading and S-representational mindreading, is 1) conceptually valid and 2) empirically robust. We are now in a position to meet the first challenge. The distinction made in section 3 between first-order and second-order mental states enabled us to conceive of relational mindreading outside of the functionalist framework of Jones's theory. Relational states may be construed as second-order states, but they need not. Thus we were able to incorporate the accounts of Ratliffe, Perner and Gordon, all of which seem to endorse a non-representational account of our default understanding of each other's goals and reasons. This already strongly suggested that the concept of relational mindreading does not depend on one's specific views about the nature of the mental states attributed. The differentiation into an intersubjective/public and a subjective/private treatment of mental representation in section 4 furthermore revealed that no concept of mental representation, functionalist or otherwise, by itself entails the subjective element of our mature understanding of belief and desire. Taken together, I contend, this goes a long way toward establishing the conceptual validity of the distinction between (the attribution of) relational and S-representational propositional attitudes. If the distinction turns out to be compatible with all influential philosophical treatments of the propositional attitude concepts, then accepting it as philosophically sound will not beg any important questions to any one of these accounts. And so, everyone should take the idea of attributing relational propositional attitudes seriously as an empirical possibility.

We have reached the turning point of this book. For now that we have met the first challenge and have established the conceptual coherence of relational mindreading, the next question is whether we should indeed expect it to be a practically robust phenomenon in folk psychological practice. This will be the topic of the next chapter.

Appendix: Ascent Routines

The discussion of relational ascent in section 3 may have reminded some readers of Gordon's *ascent routines* for propositional attitude ascription (1995, 1996, 2000b, 2007). There are two important differences, however. First, the performance of an ascent routine, unlike relational ascent, is not a form of genuine *semantic* ascent. Second, Gordon appeals to ascent routines in order to explain propositional attitude ascription within the context of his 'radical' version of the Simulation Theory (e.g. 1996). The notion of relational ascent, by contrast, has been developed independently of any simulationist considerations. This appendix briefly elaborates on these points.

Suppose someone asks you whether you believe it is raining. Rather than introspecting your mind in search of some experiential mark of belief or examining your recent behavior in light of a psychological theory, Gordon suggests that what you normally do, is simply recast the mentalistic question 'Do you believe it is raining?' as the *object-level* question 'Is it raining?' To answer *that* question, you might e.g. try to recall your last glimpse of the world outside, a few minutes ago. If the answer to this meteorological question is affirmative, you are probably prepared to *step up* a level and say 'Yes, I believe it is raining'. On Gordon's account, self-reports of mental states are accomplished by procedures "that allow one to answer a question about oneself, and specifically about one's mental states, *by answering a question that is not about oneself, nor about mental states at all…*" (2000b, 111, emphasis in original) The general idea, Gordon explains, is "that we coordinate one type of verbal behavior, self-reports of a mental state or episode, with another, the outward-looking "expression" of the state or episode." (ibid., pp. 111-112)

Gordon appeals to ascent routines in order to explain the *reliability* of self-ascriptions of mental states, and to do so without relying on introspection or invoking a folk psychological theory (see especially Gordon 2007).[76] The idea is that caregivers train children to use the correct attitudinal prefix when they express their mental states. In the case of desire, for example, they train the child to use 'I want…' when her behavior clearly indicates that she wants something, reinforcing its use by satisfying her desire when she expresses it correctly. Ascent routines do *not* explain mastery of the relevant propositional attitude concepts, however. One could teach a child who lacked the concept of belief to add the prefix 'I belief that' to her utterances whenever she makes an

76 Gordon also uses his ascent routine account to explain so-called 'Moorean paradoxes' – see footnote 62.

assertion. Although that would make her a good indicator of her own beliefs, it would not turn her into a self-ascriber of beliefs. Similarly, children who are reinforced to say e.g. 'I hope that *p*' *when* they hope that *p* thereby do not automatically self-ascribe the hope that *p*. The mere performance of an ascent routine therefore does not result in genuine *semantic* ascent, a move from expressing one's thoughts to *conceiving of one's thoughts as such* and in this sense meaningfully talking *about* them. Only someone who has already mastered the propositional attitude concepts can understand ascent routines as a form of semantic ascent.

On Gordon's radical version of the Simulation Theory, mastery of the propositional attitude concepts depends on simulation; it requires that the relevant ascent routines *be embedded within a simulation* of the person to whom the attitudes are ascribed. Recall that on Gordon's understanding of simulation, simulating another person requires that I imaginatively transform myself into the other and 'look out' onto the world from his perspective (section 2). The capacity for imaginative transformation into other 'first-persons' ensures that the ascent routine method for making *self*-reports is also applicable for ascribing propositional attitudes to *others*. As concerns belief ascription, for example, Gordon explains: "Whether in my own person or within a simulation of O, I can settle the question, "Do I believe that p?" by asking […] whether it is the case that p […] To ascribe to O a belief that p is to assert that p within the context of a simulation of O." (1995, p. 60) In general, ascription of a propositional attitude to another person consists in expression of the attitude within the context of a simulation of that person. Carrying out the relevant ascent routines then enables the interpreter to explicitly mark the simulated attitude as the attitude of the person simulated.[77]

Ascent routines thus take us from the world as simulated to the attitude of the simulated other, not from an ascribed *first-order* attitude to an ascription of a second-order attitude. On Gordon's account, comprehending other-

77 It is an interesting question how exactly one is supposed to get from *expressing* an attitude in the context of a simulation of another person to *ascribing* the attitude to the other. Gordon says very little about this. Does it require 'shifting back' to one's own egocentric frame of reference again, so as to relate the agent, now perceived from one's own point of view, to the attitudinal objects identified through simulation? Notice that ascent routines cannot provide the answer here; they only account for the reliability of (simulated) self-ascription, not for whatever psychological skills are involved in genuine other-ascription. Given this restricted role for ascent routines on Gordon's account, what is their practical use? It seems they serve the mere communicative purpose of labeling the descriptions of the world yielded through simulation *as pertaining to* the point of view of the simulated other. The need for such explicit allocation of views is especially acute when the simulated attitudes differ from one's own, as in the case of false belief ascription. Cf. Dancy's (2000, pp. 128-130) 'appositional account' of reason explanations citing the agent's beliefs (see appendix chapter 2).

ascription of propositional attitudes results in a first-order understanding of the attitude ascribed. Suppose I wonder why my colleague has still not left the office. Having recentered my egocentric map onto my colleague in imagination, I 'look out' onto the world from his perspective and use my own response mechanisms and practical reasoning skills to identify the relevant reason-constituting facts and ensuing intentions. Suppose I find myself having the simulated object-level thoughts 'it's raining', expressing the other's knowledge of the fact that it is raining, and 'let's wait a while', expressing the other's intention to wait for it to stop raining. Carrying out the relevant ascent routines then allows me to express the simulated 'it is raining' and 'let's wait a while' as the other's decision to wait a while for the reason that it is raining. The products of these ascent routines are best captured, it seems, as 'it is raining' resp. 'let's wait a while' *from his point of view,* i.e. as the first-order states 'it is raining'(O) and a 'let's wait a while'(O) (see section 3).

Note, as a final point, that Gordon's ascent routine method can be applied to both relational mindreading and S-representational mindreading. In the case of factive reason explanation, simulating the other only requires minor adjustment to one's egocentric map. Suppose I see someone running to catch a bus. The reason why she's running is the fact that the bus is about to leave. But to interpret this fact as *her* reason requires, on Gordon's story, that I approach that fact from her point of view, recentering in imagination my egocentric map onto hers. Performing an ascent routine for the implied relational state of knowledge would then allow me to say explicitly that she is aware of the fact that the bus is leaving. In the case of false belief ascription, much more drastic adjustments are needed to successfully simulate the other. In particular, it requires that I go beyond the context of the shared world, replacing the facts where necessary with the other's subjective beliefs. Yet, once the appropriate adjustments have been made, explicit ascription of belief is just a simple matter of performing the appropriate ascent routine. By itself, the ascent routine method does not distinguish between relational and representational mindreading.

On Gordon's account, the difference between relational and S-representational mindreading is a difference in the context of simulation. Relational mindreading merely requires that I simulate an egocentric change in spatial (and/or temporal) coordinates; representational mindreading additionally demands a modification of some of the facts (cf. Gordon 1992/1995). It is a matter of looking out onto the public world from the other's point of view versus adopting the other's personal, divergent view on the public world. As will become clear in the next chapter, this difference in the context of

simulation on Gordon's account applies more generally to other dominant accounts of mindreading, as a difference in the context of *interpretation*, however accomplished.

5

Making Sense in a Common World

5.1 Introduction

Chapter 3 introduced the technical notion of relational mindreading by exploiting Sellars's functionalist treatment of the propositional attitude concepts. Chapter 4 then revealed that the idea of relational mindreading is not wedded to any particular account of the propositional attitudes concepts. The general idea of 'high-level' mindreading, i.e. the attribution of propositional attitudes to agents when making sense of their thoughts and actions (see chapter 2.5), allows for both a (first- or second-order) relational and subjective (S)-representational reading. Together, chapters 3 and 4 rose to the first challenge laid out for the Relational Model of folk psychology and established the conceptual validity of the distinction between (the attribution of) relational and S-representational propositional attitudes. But is the distinction between relational mindreading and S-representational mindreading also *empirically* valid? That is: does the distinction actually make a difference in our day-to-day interactions with one another? Does it provide us with an accurate descriptive tool for modeling human discursive engagements? This was the second challenge set for the Relational Model in chapter 1, and it will be taken on in this chapter and the next. In this chapter, the focus lies on the importance

of relational mindreading in our actual daily encounters with one another. In chapter 6, attention will shift towards the complementary functions of S-representational mindreading in human discursive practice.

Thus the aim of this chapter is to show why, in addition to being a philosophically sophisticated notion, relational mindreading is also a practically robust phenomenon of human intersubjectivity. In section 2, I continue where I left off in the previous chapter. I show how the distinction between (second-order) relational mindreading and S-representational mindreading can also be made on dominant sub-personal, explanatory accounts of mindreading: the 'scientific' Theory Theory proposed by Gopnik and colleagues (e.g. Gopnik and Meltzoff 1997), the modular approach to theory of mind as put forward by Leslie and co-workers (e.g. Leslie et al. 2005), Nichols' and Stich's (2003) hybrid account, Goldman's (2006) introspectionist Simulation Theory and the so-called 'two-systems approaches' to human social cognition that have recently been developed. I conclude that the psychological plausibility of relational mindreading is not undermined by considerations regarding the nature of the cognitive processes involved. The Relational Model is neutral with respect to the cognitive implementation of mindreading; it does not preclude any dominant account found in the current literature, nor does it commit itself to any bold claims about the cognitive underpinnings of human social cognition.

Section 3 focuses on two problems that haunt the BD-Model of folk psychology: the so-called 'holism problem' and what I shall term the 'knowledge problem'. The holism problem is the problem of underdetermination of propositional attitude ascription in quotidian social contexts. I show why attempts to solve the holism problem from within the framework of the BD-Model are either psychologically implausible or question begging with respect to the Relational Model. More importantly, I show how the Relational Model *dis*solves the holism problem by taking a different perspective on the explanandum of human discursive understanding. I then direct attention to the fundamental role of knowledge attribution in the interpretation of others. The attribution of knowledge can easily be accounted for on the Relational Model, but it poses a significant challenge to the BD-Model. In short, the BD-Model seems to have no other option than to explain attribution of knowledge as the ascription of a special kind of true belief. But it is notoriously difficult to give a satisfactory analysis of knowledge in terms of true belief. And it is seriously doubtful whether any such analysis will give us a realistic model of the psychology of commonsense knowledge attribution.

In section 4 I discuss the work of several philosophers who have recently

stressed the *regulative* function of folk psychology (e.g. McGeer 2007, Hutto 2008a, Zawidzki 2008). Propositional attitude attribution, they argue, plays a crucial role in regulating each other's behavior so as to meet our interpretative needs. I place this insight in the light of the ontogeny of mindreading and re-interpret it as making a good case for the importance of relational mindreading in teaching our children the do's and don'ts of common practice. In particular, I show how Hutto's (2008a) 'Narrative Practice Hypothesis' (NPH) provides a clear example of the way in which young children could benefit from relational mindreading when taking their first steps into the space of reasons.

After having established the practical virtues of relational mindreading in our everyday encounters with one another, we should also pay attention to its shortcomings. Relational mindreading only works to the extent that we have a practical worldview in common. Section 5 explores two ways in which common practice can be extended beyond the strict confines of agential similarity. Relational mindreading allows for specification of goals and reasons relative to point of view and socio-cultural characteristics. It does not, however, have the means to account for individual differences that go beyond public agreement. The social benefits of such private understanding of other minds will be discussed in the next chapter.

5.2 Relational Mindreading on All Accounts

Many philosophers and psychologists think that our commonsense psychology defines propositional attitudes as representational mental states. This has generally been taken to imply that our ordinary understanding of one another as being intentionally directed at propositionally articulated goals and reasons is a form of mature belief-desire psychology. The previous chapter revealed that this does not follow. An understanding of the mental representation relation *simpliciter* does not entail a subjective understanding of what is represented. Once this is appreciated, it is possible to tease out the difference between relational and S-representational mindreading even on accounts that explicitly adopt the BD-Model. Importantly, most of these accounts have been presented in the literature as *explanatory* accounts of mindreading, directed at the sub-personal level of social cognition. The Relational Model, by contrast, is primarily a descriptive account directed at an important explanandum of human intersubjectivity. By isolating a relational conception of the explanatory posits of mentioned sub-personal accounts, however, it becomes vividly

clear that the Relational Model of human discursive engagement is prima facie compatible with *all* existing accounts of goal-reason attribution in the current literature, whether targeted at the personal, descriptive or the sub-personal, explanatory level of social cognition. And this, in turn, should erase any remaining skepticism concerning the cognitive feasibility of relational mindreading.

Our discussion of Sellars's Myth of Jones in chapter 3 has already opened the door to an alternative interpretation of the Theory Theory, according to which the central posits in our folk psychological theory are relational mental states. As we have seen, this interpretation also throws a different light on Lewis's (1972) influential account of analytic functionalism, an account which adopted "the working hypothesis that [Sellars's] myth is a good myth", i.e. that "our names of mental states do in fact mean just what they would mean if the myth were true." (p. 213) Armed with the distinctions made chapter 4.3, we could say that TT naturally invites a second-order reading of relational mindreading: the posits of our folk psychological theory are FRR-states, the occupants of functional roles defined in terms of perceptual input (entry transitions), behavioral output (exit transitions) and each other (inferential transitions).

The shift towards S-representational mindreading would then consist in adding a subjective dimension to the functionally classified world-directed attitudes attributed. Along Lewisian lines, we could speculate that this is accomplished by imaginatively placing the world-directed attitudes attributed in an 'near possible world', such that the functional classifications of the states attributed to the agent in the *actual* world mirror the entry/inference/exit transitions of the *relational* states the agent would (and ought to) have had in this *possible* world. Accordingly, ascription of the false belief that *p* to an agent would require that one conceive of the agent's actual state as a relational attitude of belief in a near possible world in which *p*, and then 'project' this possible world onto the actual world in which the belief that *p* is ascribed, yielding the agent's personal view on the world, an 'agent-centered possible world' (Quine 1969; Lewis 1979) according to which *p*.[78]

78 This comes close to the idea Dennett (1987, ch. 5) expressed in terms of a 'notional attitude psychology', a variety of the intentional stance according to which we attribute 'notional attitudes' to one another, attitudes that relate an agent to a 'notional world'. The idea of a notional world is the idea of an agent-centered fictional world, of "a model […] of one's internal representations." This model, Dennett explains, "*does not consist itself of representations but of represenetds*. It is the world "I live in," not the world of representations in me." (p. 154, emphasis in original) On this proposal, the shift to representational mindreading would consist in a shift from attributing strictly relational attitudes in interpreting someone's world-directed behavior, to attributing mere 'notional' attitudes.

Thus we can construe a TT-account of belief-desire psychology on which the folk psychological theory *itself* is *entirely* framed in (second-order) relational terms, and on which overcoming the informational constraint of relational mindreading is explained in terms of the *application* of the relational theory in a counterfactual context of interpretation. Addition of the required subjective dimension for belief-desire psychology would consist in applying one's relational theory of mind to imagined possible scenarios the interpreter projects onto the actual situation of the interpreted agent. But on such construal, one would expect that the relational theory could often simply be applied 'extensionally' in the actual world, i.e. without taking into account the 'agent-centered possible world' that informs the agent's behavior in the actual world. Interpretation of other people's world-directed attitudes would be strictly relational by default; agent-centered possible scenarios would be called upon especially when relational interpretation runs aground.

It seems these considerations directly apply to the empirist, 'scientific' versions of TT. Recall from chapter 2.3 that scientific TT holds that our folk psychological theory is acquired in much the same way as real scientific theories develop, i.e. by means of theory construction and revision, underpinned by domain general learning mechanisms rather than innate modules devoted to mindreading. Our alleged folk psychological theory is also thought to show important structural similarities to real cognitive-scientific theory, consisting of tacitly but explicitly represented knowledge of the causal generalizations linking mental states to input, output and other mental states (e.g. Gopnik and Wellman 1992; 1994; Gopnik and Meltzoff 1997). As indicated in chapter 3.5, this proposal invites an interpretation according to which this quasi-scientific folk psychological theory is an explicit version of the theory Jones taught the Ryleans. On this reading, the theoretical posits of our folk psychological theory would be FRR-states. Gopnik et al. are not specific about the nature of the representational states that are supposed to figure on our mature representational theory of mind. But whatever notion of mental representation is involved, it seems a perfectly coherent option that *qua* functionally defined theoretical concept, it can be deployed without addition of the subjective dimension required for representational mindreading.

In the debate on mindreading, it is often argued that our default interpretation strategy is to attribute our own beliefs to others. Leslie et al. (2005) are particularly clear about this:

> A true-belief default is ecologically valid because, at least about mundane matters, people's beliefs usually *are* true. We can go a little further than

> this. For a basic belief-attributing system—one whose business concerns simple everyday beliefs—the true-belief attribution *ought* to be the default. This is because, in the absence of specific information, the only general constraint on belief attribution is provided by the state of the world (as it appears to the attributer). (pp. 48-49, emphasis in original)

In similar fashion, Nichols and Stich (2003) point out that "people are very good at attributing beliefs to others even in cases where they have no apparent evidence." (p. 69). They explain this in terms of a process of 'default belief attribution' in which the interpreter lets her own beliefs enter into her model of the interpretative target's beliefs. Now, before having a closer look at Leslie's and Nichols and Stich's respective proposals, it should be noted that we could easily explain default true belief attribution in terms of default attribution of the *relational* attitude of 'belief'.[79] In fact, Leslie's et al. remark in the quote above, that "in the absence of specific information, the only general constraint on belief attribution is provided by the state of the world," readily suggests a relational understanding of the default attribution of shared beliefs. If, in the absence of specific information, we are supposed to interpret other people's behavior in terms of the state of the world, then why couldn't the beliefs we attribute by default be states that simply relate the agent to the state of the world?

On Leslie's modular account of mindreading (see also Leslie and Thaiss 1992, Leslie 1994, Leslie and Polizzi 1998, Scholl and Leslie 1999, Leslie et al. 2004), our mindreading abilities are subserved by two central mechanisms: an innate modular 'theory of mind mechanism' (ToMM) and a non-modular 'selection processor' (SP), which develops through first few years of life. ToMM incorporates innate concepts of propositional attitudes such as belief and desire and its job is to spontaneously infer form observed behavior candidate contents for the mental states to be attributed to the agent. SP is supposed to select among competing candidates those contents that best fit all the evidence, including e.g. the agent's past behavior, his whereabouts, etc. As Scholl and Leslie (1999) make clear, "ToMM always makes the current situation available as a possible and even preferred content, because (a) the current situation is a truer picture of the world, and (b) beliefs tend to be true." (p. 147) Ascribing false beliefs "requires this default interpretation to be inhibited by SP, so that the weaker [less salient] false content be selected." (*ibid.*)

79 That is: a state of common knowledge. See section 5.3.

Now why should we assume that the contents that ToMM generates always include counterfactual contents, contents that are incompatible with the interpreter's own beliefs? In many standard social situations, it seems, there are simply no such counterfactual possibilities available to be automatically inferred by ToMM. Consider seeing someone hiding for the rain under a tree. He hides under the tree because (he believes) it is raining. Even when we consciously think about the situation, it may be hard to come up with an alternative explanation as to why he stands under the tree. It seems implausible that, unconsciously, ToMM actually provides several. I see no reason, then, to concur with Leslie et al. (2005) that, according to their own account, "The *typical* mode of operation (MO) of ToMM is to offer more than one candidate content for a mental state attribution." (p. 49, emphasis added) And even if ToMM does infer more than one candidate content for the belief to be attributed, these contents will often be reflecting the real situation as assessed by the interpreter. Perhaps the man is hiding *behind* the tree because he believes the two policemen over there are looking for him. Here we have two competing contents for the belief to be attributed in order to explain the man's behavior. But why couldn't SP here be selecting between two *relational* 'beliefs'? ToMM is supposed to be a 'metarepresentational' device (Leslie 1987, 1994), but as we have seen in chapter 4.4, metarepresentational interpretation does not automatically amount to S-representational mindreading. Whatever concept of representation ToMM is supposed to incorporate, it seems it can be deployed without the subjective dimension required for S-representational mindreading.

One could speculate that ToMM operates under the disjunctive constraint outlined earlier (see chapter 4.4), inferring *either* world-directed states with content that reflects the real situation *or* suppositional states with counterfactual contents. It would then be SP's job to inhibit, in false belief conditions, true contents and select the counterfactual contents as the contents of the target's world-directed attitudes, thus marking it as the target's subjective representational state. When ToMM does not pass counterfactual contents on to SP, the latter would issue a mere relational understanding of the interpreted agent.[80]

On Nichols and Stich's account of third-person mindreading (2003), the

80 Cf. Doherty's (2009, pp. 51-54) discussion of Leslie's ToMM/SP account. He suggests a reading on which ToMM only comes with the concept of 'prelief' (Perner et al. 1994, Perner 1995), leaving full-blown 'theory of mind' reasoning to SP. As indicated in section 4.3, interpreting others in terms of 'preliefs' would be an instance of relational mindreading, also subject to the disjunctive constraint.

mindreader builds a model of the interpretative target in what they call her 'possible world box' (PWB), a workspace "in which our cognitive system builds and temporarily stores representations of one or another possible world." (p. 28) The PWB, in other words, is a device for hypothetical reasoning, using the same inference mechanisms that are used for the formation of real beliefs (cf. p. 85). In order to explain default belief attribution, they hypothesize that "when the PWB gets co-opted for mindreading, all of the mindreader's own beliefs are included in the model of the target's beliefs that is being built in the PWB." (p. 85)

The first thing to notice here is that this explanation of default belief attribution appears to be slightly at odds with their own characterization of the PWB as containing token representations whose job "is not to represent the world as it is or as we'd like it to be, but rather to represent what the world would be like *given some set of assumptions that we neither believe to be true nor want to be true*." (p. 28, emphasis added) For default belief attribution ensures that the model of the target's beliefs *is* a model of the world as the interpreter believes it to be. What Nichols and Stich mean, is that only the *contents* of the mindreader's belief are by default transferred to the PWB, dissociating them from their normal functional role in guiding the interpreter through the world. Still, that does not imply that in the case of default belief attribution, the model of the target should be anything other than a model of the *actual* world, i.e. a copy of the interpreter's own assessment of the world. There appears to be no reason why this model of the interpretative target's belief should be marked as a *possible* world considered by the target *as actual*. It would seem to suffice to merely mark it as the target's model of the actual world. If so, our default belief attribution strategy comes close to the idea expressed in chapter 4.4, i.e. of conceiving of other people's minds as models, pictures or maps of reality. But as we have seen, merely adding a pictorial intermediary to the target's relation to the world does not amount to an appreciation of the target's subjective view on the world. The beliefs attributed by means of default belief attribution would be relational 'beliefs', not subjective representational beliefs.

So in the case of default belief attribution, why couldn't the interpreter simply build a model of the target's belief in her 'actual world box', i.e. her 'belief box'? She would simply use her own inference mechanisms on her own beliefs and let the contents enter into the model of the target's beliefs. Crucially, the model would not be marked as a model of a possible, and possibly counterfactual, world. Nichols and Stich's PWB would be left out the

interpretative loop in these standard cases of default belief attribution.[81] The PWB would only be called upon for belief attribution when and insofar as 'discrepant belief attribution mechanisms' (p. 87 ff.) kick in. When discrepant beliefs are detected, the PWB could work out the consequences of those beliefs and feed those back into the default model of the target in the interpreter's belief box. The interpreter's inference mechanisms would then ensure that the model of the target's beliefs remains consistent and that some of the interpreter's own beliefs copied into the model of the target get 'erased' and supplanted by the incompatible beliefs worked out in the PWB. The interpreter's model of the target's beliefs would be a model of the actual world modified only to the relevant extent in order to allow for discrepant beliefs. To that extent, the model of the target's beliefs would be marked as a model of a counterfactual world, considered by the target as actual, i.e. as the target's subjective beliefs of the world.

On both Leslie's and Nichols and Stich's account, then, we can understand default belief attribution as consisting in the attribution of relational beliefs, and hence as an instance of relational mindreading. The conclusions we drew from Sellars's myth of Jones thus not only apply to strictly theoretical renderings of FP, such as scientific TT. Also on the modular account of Leslie et al., as well as the hybrid account of Nichols and Stich, a relational dimension of mindreading is easily incorporated.

The same goes for Goldman's account of the Simulation Theory (e.g. 2006). On Goldman's account, third-person action prediction proceeds by 1) pretending to be in the other person's situation, i.e. 'putting oneself in the other's shoes', 2) using one's own resources for practical reasoning in order to figure out what one would do in that situation, 3) introspectively classifying the pretend-output as one's intention or decision to perform a certain action in the pretend scenario, and finally 4) attributing that decision in non-pretend mode to the interpretative target. The simulationist aspect of Goldman's proposal lies in the central role played by the interpreter's own practical reasoning system, fed by pretend beliefs and desires. However, Goldman allows for elements of TT to enter into the overall story. Theoretical resources may be required to figure out the right kind of pretend input, i.e. which pretend beliefs and desires to feed into one's own practical reasoning system.

But making such theoretical inferences would be unnecessary in standard cases in which one simulates under the assumption that other people share one's own beliefs and desires. In these cases of 'default belief and desire at-

81 Cf. Nichols and Stich (2003), figure 3.5 and 3.6 (pp. 93-94).

tribution', a description of the actual situation of the target would suffice: one could simply feed that description into one's offline practical reasoning system and see what happens. On Goldman's story, the outputs generated by the interpreter's offline practical reasoning system are classified as mental state kinds by means of introspection. Thus, in the case of action prediction, one would find oneself in the pretend scenario thinking to oneself 'let's do *a*!' and subsequently classifying this as one's pretend-decision to perform action *a*. Then one performs an analogical inference from oneself to the other and judges that the other will decide to perform action *a*. As long as the interpreter simulates under the basic assumption of a shared take on the world, it seems unnecessary that anywhere along the simulation route, the interpreter conceives of her own mental states as subjective representational states. The pretend decision can simply be classified as a (second-order) relational state of 'being in a pretend state of intending to do *a*'. Goldman thinks that introspection proceeds by recognizing an 'introspective code' or 'I-code' (2006, pp. 260-264) that is attached to the pretend-output generated by the practical reasoning system. He identifies three possible dimensions or parameters the I-codes might have: the doxastic/credal dimension, the preference or valence dimension and the bodily feeling dimension (p. 261). None of these dimensions include the particularly *subjective* dimension that would be involved in first-person S-representational mindreading. In other words, one could easily maintain, on Goldman's story, that the interpreter observes herself simply as being related to a (pretend) situation, rather than as having a particular or peculiar view on the (pretend) situation. Through analogical inference, it would by such relational states that are attributed to the simulated agent.

The same line of reasoning can be applied to Goldman's story about explanation of actions already preformed in terms of beliefs and desires. Goldman dismisses the idea that the practical reasoning system can work backwards, from intentions to the beliefs and desires that generated it. Instead, he proposes to approach this kind of 'retrodictive' mindreading via a generate-and-test strategy (2006, pp. 183-185). On Goldman's introspectionist version of ST, this is supposed to work as follows. The interpreter asks herself why she would perform or have performed the particular action under consideration, were she to be or have been in the target's situation. She then generates certain possible candidate pretend states and runs these through her offline practical reasoning system. If the resulting pretend decision matches the observed action of the target, she attributes the generated states to the target, again through introspective classification and analogical inference. If the pretend decision doesn't match, she tries again with other candidate states, until she hits upon an out-

put that does match. Goldman suggests that the hypothesis generation part of the 'generate and test strategy' might be mediated by theory, perhaps with the help of prior simulations. The test part of the strategy would essentially be a simulation process.

Again, we should realize that in default cases of action explanation, the pretend situation generated to run one's simulation on may simply mirror the actual situation of the world (at present or in the recent past) according to the interpreter's own assessment. The states attributed upon finding a match could be classified as relational states of belief and desire, which upon attribution to the target would be conceived as relating the agent to her goals and reasons for action. Even if one wishes to hold, as Goldman seems to do, that hypothesis generation is mediated by theory, we should allow for the possibility that the states issued by the theory are (second-order) relational states. So it seems that, also on Goldman's introspectionist Simulation Theory, there is room for a relational dimension of third-person mindreading.

As a last illustration of the feasibility of relational mindreading at the subpersonal, explanatory level, consider the so-called 'two-systems approaches' to human social cognition that have recently been proposed in the literature (e.g. Carruthers 2006, Goldman 2006, Apperly and Butterfill 2009, Apperly 2011). Many theorists these days make a distinction between low-level social cognition and high-level social cognition. As explained in chapter 2.5, the distinction between low- and high-level can be applied to 1) the nature of the cognitive process underlying the socio-cognitive activity (fast, efficient, inflexible, involuntary and subconscious versus slow, effortful, flexible, voluntary and conscious) and 2) to the nature of the states tracked by that activity (simple, observable states tracking specific types of behavior versus sophisticated, unobservable states with only tenuous, holistic connections to behavioral types). The general idea of two-systems approaches is that our socio-cognitive abilities are underpinned by two separate cognitive systems: one for low-level and one for high-level social cognition. On most proposals, the system that operates at a low, i.e. fast, efficient, inflexible, etc. level is also the system dedicated to tracking low-level states. Here we can think of the detection of simple goal-directed behavior (e.g. Gergely and Csibra 2003), face-based emotion recognition (e.g. Goldman 2006) or the tracking of non-propositional, belief-like states (Apperly and Butterfill 2009). The system that operates at the high, i.e. slow, effortful, flexible, etc. level is required for interpretation in terms of high-level states, paradigmatically the propositional attitudes. This two systems approach for social cognition is all but *ad hoc* from the point of view of cognitive psychology and cognitive neuroscience. It has been successfully applied to

explain performance in other cognitive domains, e.g. on number cognition tasks (see Apperly and Butterfill 2009 for an overview).

The point of bringing this up is that one may be tempted into thinking that, according the abovementioned hypothesis, the Relational model should now postulate *three* systems: one low-level system for tracking low-level states and *two* high-level systems for tracking relational propositional attitudes and S-representational attitudes, respectively. This is a rather bold empirical claim about the sub-personal implementation basis of human social cognition, a claim that certainly deserves to be met with a healthy dose of skepticism. And as far as I know, there is no empirical support from other domains of human cognition for such 'three systems approach.' From this perspective, then, the Relational Model is empirically unattractive. Better stick to *one* system for high-level social cognition and disregard the conceptual distinction between relational and S-representational high-level mindreading as psychologically or cognitively invalid.

This worry about cognitive implementation would be well grounded *if* the Relational Model indeed implied two separate cognitive systems for propositional attitude attribution. But it does not. As I explained in chapter 4.3, the psychological difference between relational and S-representational mindreading lies in the information available for interpreting another person's world-directed attitudes. S-representational mindreading takes away the disjunctive constraint on relational mindreading and allows information incompatible with the interpreter's own assessment and appraisal of the world to figure in the content clauses of the world-directed attitudes attributed to others. The switch from relational to S-representational interpretation makes a substantial difference from a *folk-psychological* point of view (see chapter 6). In the present context, however, it is worth stressing that as a *cognitive difference*, it does not invite any grand claims about sub-personal architecture. Quite plausibly relational mindreading and S-representational mindreading use much the same cognitive resources, whatever they turn out to be. The only and crucial difference is that S-representational mindreading *additionally* requires the capacity to inhibit one's own inferentially articulated assessment of the world from interfering with the contents of the world-directed attitudes ascribed to others. As we have seen in this section, this additional capacity has been cashed out in various ways at the sub-personal level, e.g. in terms of recruitment of Leslie et al.'s 'selection processor' or Nichols and Stich's 'possible world box'. These and other sub-personal accounts of high-level mindreading that have surfaced in the literature are specifically designed for the task of S-representational mindreading. This should come as no surprise, since they are all inspired by the

BD-model. The point to take notice of is that the Relational Model claims that mindreading often involves less than that. Whatever the nature of cognitive processes that subserve the abovementioned capacity for content inhibition, the Relational Model predicts that these processes are *supplementary* to our basic capacity for propositional attitude ascription and are often left out of the cognitive loop. The Relational Model does not need to postulate anything in addition to the sub-personal mechanisms that have been invoked to explain belief-desire ascription. Worries about the cognitive implementation of relational mindreading are ill-founded.

5.3 Epistemic Holism and Default Knowledge Attribution

In chapter 2.6 I explained in what sense propositional attitudes should be considered as 'unobservable' states. Propositional attitudes are unobservable in the sense that they only have tenuous connections to specific behavioral types. This, I indicated, has everything to do with the holistic nature of propositional attitude ascription.

Ever since the downfall of logical behaviorism, it has been widely recognized that no particular piece of behavior can be seen as evidence of the presence of any particular (cluster of) propositional attitude(s), and that no particular (set of) propositional attitude(s) can be regarded as issuing any particular kind of behavior, without taking into account an inferentially articulated mental context in which the ascribed attitude(s) make proper sense in relation to the behavior in question. As explained in chapter 2.6, the relation between propositional states and their behavioral manifestations is one-to-many and many-to-one. Interpreting John's walking down the street in terms of the reason that he has run out of milk, i.e. as being informed by a state with the content that he has run out of milk, only makes sense if we are able to place that state in the wider, temporally extended context of John's mind: his further knowledge, intentions, plans, preferences, etc. Awareness of the fact that one has run out of milk, devoid of any further mental context, does not suggest any particular kind of behavior, and so every kind is a possibility. The things that go on in the world on the one hand, and an agent's behavioral responses to those things on the other, are not reliably correlated with individual (sets of) propositional attitudes, only with indefinitely large, whole systems of such attitudes. As Davidson (1970/2001a) observes:

> There is no assigning beliefs to a person one by one on the basis

> of his verbal behaviour, his choices, or other local signs no matter how plain and evident, for we make sense of particular beliefs only as they cohere with other beliefs, with preference, with intentions, hopes, fears, expectations and the rest. (p. 221)[82]

The holistic nature of propositional attitude attribution has often been presented as an epistemic *problem* that we have to face in our quotidian discursive engagements with one another (Morton 1996, 2003, Bermudez 2003, 2009, Zawidzki 2008, Apperly 2011). Zawidzki (2008) states the problem thusly: "Holism is a problem because it leads to underdetermination: any finite set of behavioral evidence is compatible with an infinite number of distinct sets of propositional attitudes. In addition, ascription of any finite set of propositional attitudes is compatible with an infinite number of distinct behavioral predictions." (p. 196) In other words, since interpretation of other people's actions is always based on finite behavioral evidence, there is no easy way of telling which specific set of propositional attitudes best explains the behavior in question. And since any ascribed, finite set of propositional attitudes can be made compatible with indefinitely many distinct behavioral predictions (as long as the appropriate adjustments are made to the agent's background mental states) there is no straightforward way of knowing how the agent will behave in the future on the basis of the ascribed set. Mindreading almost starts to look impossible (cf. Apperly 2011, pp. 118-119). So how do we manage to attribute goals and reasons to each other in daily social life? And how do we do this, moreover, in a way that is fast and reliable enough to be of genuine practical use?

The problem is particularly acute for Theory Theory accounts of mindreading. Recall that according to TT, providing folk psychological explanations and predictions is to subsume observable behavior under general principles that spell out how behavior and environmental conditions link and give rise to mental states, how mental states relate and give rise to one another, and how they cause particular kinds of behavior. The problem is that

> we can only apply these principles if we can identify, among a range of possible principles that might apply, the ones that are most salient in

82 As is well known, Davidson took the *epistemic* holism of propositional attitude ascription to imply the semantic holism of the contents of the attitudes ascribed. Issues regarding semantic holism run orthogonal to our present concerns, however. Putting questions regarding the indeterminacy of propositional attitudes to one side, I will therefore only speak of the *under*determination of propositional attitude ascription.

> a given situation. We need to identify whether the appropriate background conditions hold, or whether there are countervailing factors in play. We need to think through the implications of the principles one does choose to apply in order to extrapolate their explanatory/predictive consequences. The need to do all these things makes folk psychological generalizations rather unwieldy. (Bermudez 2009, p. 194)

In order to tackle the problem of epistemic holism, our folk psychological theory would have to come accompanied with volumes of ceteris paribus clauses telling us how the theory should be applied in the particular context of interpretation. Successful discursive interaction would furthermore require that we sift through all these volumes on the spot. The underdetermination of propositional attitudes by behavioral evidence threatens to turn the task of accurate propositional attitude ascription into an intractable search problem (cf. Zawidzki 2008).

There are actually two worries here. The first concerns *computational complexity*: it is the worry voiced by Bermudez that applying folk-psychological principles to specific situations requires computation over a host of additional clauses to determine whether the appropriate background conditions hold and whether there are countervailing factors in play. Another closely related but deeper worry is that TT here faces a precursor to the frame problem: the problem of *determining relevance* (cf. Heal 1996, Wilkerson 2001). As Spaulding (2010, pp. 136) puts it, the problem is "how one determines which general principle to apply in a particular case given that the relevant information for determining which principle is appropriate is in principle unlimited and could come from any domain."

It is mainly for these reasons that there are no self-acknowledged 'pure' theory theorists in the debate. Pure or 'strong' TT, as Heal (1998) terms it, holds that our knowledge about other people's minds is arrived at independently from our knowledge about the world around us. In specific, it claims that our ability to determine the contents of the mental states of others functions independently from our ability to form thoughts with those same contents ourselves. On this account, thinking about some particular subject matter is one thing, thinking about other people's thoughts about that subject matter another thing entirely. This strong version of TT is committed to the claim that we come equipped with a (tacit) folk theory of content, a theory that specifies the contents of other minds in strictly causal terms *without* drawing from our own response mechanisms and reasoning skills.

Contrast this pure account of TT with Sellars's version in his Myth of

Jones. As explained in chapter 3.5, Jones taught the Ryleans his new theory by exploiting their pre-existent linguistic know-how. He did *not* teach them a theory of meaning. We found no reason to think that the Ryleans' interpretation of each other's overt verbal behavior itself depended on a functionalist theory of meaning, a theory that explicitly states the transition rules that specify the functional roles of the expression in their language. As far as the Ryleans were concerned, I argued, their functional role semantics was not a theory that *explained* the meaningfulness of their utterances; it was only a device that *classified* their already meaningful utterances in a functionalist way, so as to enable them to comment upon and criticize each other's linguistic performances. These functional classifications were parasitic on a prior, non-theoretical understanding of each other's verbal behavior. It was this classificatory apparatus that Jones used to characterize the contents of the theoretical posits of his new theory, the inner episodes he calls 'thoughts'. Once the Ryleans had become skilled Jonesian mindreaders, they could determine the contents of the mental states they ascribed by exploiting their non-theoretical understanding of the meaning of each other's linguistic utterances.

Sellars left enough room for his Ryleans to determine the contents of other minds by drawing from their own recognition and reasoning skills. And so, it seems did TT-ists after Sellars. According to Nichols and Stich (1998), Heal's 'strong' theory-theorist "is a straw man, a figment of Heal's imagination." Heal contrasts strong TT with her version of the Simulation Theory in terms of 'co-cognition', which is "just a fancy name for the everyday notion of thinking about the same subject matter […] Those who co-cognize exercise the same underlying multifaceted ability to deal with some subject matter." (1998, p. 483) Heal's alternative to strong TT is the claim that:

> It is an a prior truth that thinking about others' thoughts requires us, in the usual and central cases, to think about the states of affairs which are the subject matter of those thoughts, i.e. to co-cognize with the person whose thoughts we seek to grasp. (1998, p. 484)[83]

83 Regarding the notion of a priori she uses here, Heal states that: "To say that something is a priori, as I mean it here, is not to say that it is susceptible of proof in some formal system. The idea is rather that the a priori is that which is deeply embedded in our world view […] An a priori claim is one we rely on unhesitatingly in making inferences; in cases where it seems threatened our automatic assumption is that the threat is illusory and we seek was of explaining it away; if challenged we are thoroughly at a loss to describe realistically or in any detail how we would carry on intellectually if we could not rely on it. Hence the a priori is not something the testing of which could be an object of a realistic scientific project. To say that a judgement is a priori in this sense is not to say that it will never be abandoned or replaced; nor is it to say that we know that the concepts invoked in it could not mutate into what are recognizably successors in terms of which

On the most plausible interpretation, this claim reads that "in thinking about others' thoughts, we typically use co-cognition and if we could not rely on it, the ordinary mindreading that we employ in our daily interaction with one another would be severely disrupted." (Nichols and Stich 1998, p. 508) Nichols and Stich, however, find this interpretation of the co-cognition thesis "so obvious that [they] wonder who is supposed to disagree with it." (p. 509) Heal's a priori claim about the role of co-cognition in everyday interpretation, they conclude, 'turns out to be a banal truth that no one has ever questioned." (p. 511)

So if pure TT is not an option, what are the alternatives to counter the problem of epistemic holism? The answer that is 'so obvious' according to Nichols and Stich is that we use our capacity to co-cognize with others to determine which goals and reasons most likely figure as the contents of their mental states when interpreting their actions. We have already seen several proposals as to how this is supposed to be accomplished in the previous section. Nichols and Stich hypothesize that we build a model of the interpretative target in our 'possible world box' (PWB), a general device for hypothetical reasoning that exploits the same inference mechanisms that are used for the formation of real beliefs. When the PWB is co-opted for mindreading, our own beliefs are by default included in the model of the target's mind. As we have seen, this is how they account for 'default true belief attribution'. Goldman's story is not very different in this respect. He too claims that we tend to let the contents of our own mental states enter into the content clauses of the pretend-states that represent the interpretative target's mind, and which are subsequently fed into our own practical reasoning mechanisms.

What is not so obvious, however, is why this solution to the holism problem should be framed in terms of the BD-Model. If fast and reliable mindreading in our daily discursive encounters with one another requires that we by default rely on our own understanding of the world around us, then why should the contents of the minds of others thusly determined always be processed in an S-representational way? Given that we have no other option than to rely on our common understanding of the world, why have us explicitly mark the contents assigned to each other as giving expression to how *the other in particular* represents the world, a representation which only as a matter of contingent fact matches the way we ourselves take the world to be? As long

the claim is false. But it is to say that at the moment we have no serious idea about how such replacement or mutation might go and hence little powerful argumentative work can be done by invoking such shadowy and perhaps illusory possibilities." (Heal 1998, p. 480)

as the people we interpret in our day-to-day encounters think and act in line with the way one is supposed to think and act in their circumstances, adding this S-representational dimension is simply unnecessary.

The Relational Model moreover provides a clear alternative characterization of the interpretation process. And what is more important: the Relational Model *implies* the solution to the holism problem suggested by the accounts mentioned above. At the relational level of discursive understanding, interpreters have no other option than to rely on their own inferentially articulated appreciation of the world around them in the course of making sense of the thoughts and actions of others. We could also say that the Relational Model makes the *problem* of epistemic holism disappear. It effectively dissolves the problem by giving a different characterization of the standard situation faced by folk psychological interpreters. Epistemic holism of propositional attitude attribution only appears as an epistemic problem as long as our discursive engagements with one another are portrayed as attempts to gain access to each other's private and potentially discrepant understanding of the world. The Relational Model, by contrast, claims that most of our discursive engagements with one another take place on common ground, relating each other's actions to salient worldly features that solicit a response, and thinking about each other's thoughts by looking out into the world and letting ourselves be guided by the inferences that it licenses. As long as most of our propositional attitudes are bound by our common understanding of things, the epistemic holism of propositional attitude attribution can be just as (un)problematic as the holism of propositional attitude instantiation. Thinking about thinking need not create any higher degree of epistemic uncertainty than thinking itself.

Against this, it might be objected that I overestimate the similarities between different agents, and thereby overestimate the practical use of relying on a shared understanding of the world, i.e. on relational mindreading in everyday contexts of interpretation. Relational mindreading only works to the extent that interpreter and interpretee share their view on the world, have the same preferences, make the same priorities, etc. But this is an assumption that cannot be taken for granted. Differences in cultural and religious background, class, gender, profession, character and personality are profound. Given such interpersonal differences, relational mindreading will not get us very far in daily social interaction. It will often fail as a means to make sense of other people's actions or to provide accurate predictions of their future behavior. It is hard to believe that we rely on a form of mindreading with such severe restrictions as our primary strategy for attributing goals and reasons to others in daily social life. Or so the argument goes.

In response to this worry, let me here simply make the observation that it is much harder to direct attention to all the things we have in common than to focus on our differences. The extent to which relational mindreading is deemed insufficient for daily social practice is easily overstated. It is mostly in situations in which we do *not* share a common appreciation and understanding of the world that explicit use of S-representational belief-desire psychology is called upon in order to secure successful coordination and cooperation in daily social life (see chapter 6). But that should not lure us into thinking that this is the default situation we find ourselves in during our discursive interactions with others. The fact that we speak the same language and are able to successfully communicate with one another should suffice as an indication of the vastness of the common ground we stand on. In Davidson's words, "To understand the speech of another, I must be able to think of the same things she does; I must share her world." (1982/2001c, p. 105) Communication, he continues, "depends on each communicator having, and correctly thinking that the other has, the concept of a shared world, an intersubjective world." (*ibid.*) What the above considerations about epistemic holism show, is that successfully conversing with one another cannot get off the ground without approaching one another as having a shared, inferentially articulated background view on the way things are or how they should be. There is more to say about this objection, in particular about the assumption behind it, but that will have to wait until section 5.5.

There is another problem with the proposed solution of 'default true belief attribution.' This is what I call the 'knowledge problem'. In the debate on folk psychology, the attribution of knowledge, as opposed to mere true belief, has generally been ignored (but see Gordon 1987, Hornsby 2008, Perner and Roessler 2010, see also Williamson 2000). Yet simple reflection reveals that default action explanations involve more than mere true belief ascription. Consider the following example provided by Hornsby (2008). The example concerns

> Edmund who believes that the ice in the middle of the pond is dangerously thin, having been told so by a normally reliable friend, and who accordingly keeps to the edge. But Edmund's friend didn't want Edmund to skate in the middle of the pond (never mind why), so that he had told Edmund that the ice there was thin despite having no view about whether or not it actually was thin. Edmund, then, did not keep to the edge because the ice in the middle was thin. Suppose now that, as it happened, the ice in the mid-

> dle of the pond was thin. This makes no difference. Edmund still didn't keep to the edge because the ice was thin. The fact that the ice was thin does not explain Edmund's acting, even though Edmund did believe that it was thin, and even though the fact that it was thin actually was a reason for him to stay at the edge. (p. 255)

Hornsby uses this example to show that the 'factive' explanations of action we provide in our everyday lives (cf. Gordon 2000a, 2001), such as 'Edmund kept to the edge of the pond because the ice in the middle of the pong was thin,' imply the attribution of knowledge, in this example: that Edmund knows that the ice in the middle of the pond is thin. As Hornsby goes on to explain, such factive, knowledge implying explanation would be infelicitous in the above example. Edmund didn't keep to the edge *because* the ice in the middle of the pond was thin. That could not have been *his* reason for keeping to the edge; he did not have the right sort of epistemic connection to that reason. Edmund is a familiar sort of character in epistemology, of the type originally designed by Gettier (1963), to show that justified true belief is not sufficient for knowledge. Most of our commonsense explanations of action, such as 'John went to the supermarket because his fridge was empty' or 'John walked down the road in order to buy some milk at the supermarket' imply more than just justified true belief – they imply knowledge: that John *knows* that his fridge is empty, that he knows that the supermarket is somewhere down the road, etc.

For our purposes, the essence of Gettier cases concerns the interpretative act on the basis of which we judge the protagonists to lack knowledge about the subject matter their beliefs are about. In order to assess Gettier cases, we have to adopt an S-representational interpretative stance. We assess the protagonist's belief as being *justified* from *his* subjective point of view, whereas we judge his belief as being *true* from *our own* (subjective) point of view. From our point of view, the truth of the belief is a matter of sheer epistemic luck. We judge the protagonist to lack the right sort of epistemic contact with that which his belief is about in order to be granted a state of genuine knowledge.

The knowledge problem that confronts the BD-Model is how to explain the kind of implicit knowledge attribution that subserves our everyday reason explanations in terms of subjective belief ascriptions. The problem is how to account for the 'right sort' of epistemic contact that knowledge implies, as judged from *our* perspective as interpreters, in terms of the epistemic properties of the agent, as judged from *his* subjective perspective. For attribution of genuine knowledge, the agent should somehow be able to reach out of his private 'centered possible world' and make contact with the world as we

as interpreters experience it. Gettier cases clearly demonstrate the difficulty of giving a satisfactory analysis of our implicit folk concept of knowledge in terms of subjective belief. But if epistemology has taught us anything over the last 50 years, it is that there is no easy way to account for default knowledge in terms of some form of subjective belief. And it is seriously doubtful whether any such analysis is forthcoming. Yet the BD-Model seems to have no other option. What is more, any proposed analysis should also meet certain criteria of psychological plausibility. For as far as the BD-Model is concerned, such analysis would not just be a philosopher's sophisticated understanding of the concept of knowledge, it would actually have to be applied in the context of our quotidian, fast and flexible interactions with one another.

As with the holism problem, the Relational Model effectively dissolves the problem of default knowledge attribution. For it holds that implicit knowledge attributions that subserve our commonsense understanding of everyday reason explanations are a form of relational, rather than S-representational mind-reading. It holds that default 'factive' reason attribution involves ascription of relational states of 'belief', i.e. relational doxastic states with a mind-to-world direction of fit (see chapter 2.4). The Relational Model avoids the problem presented by Gettier cases because it doesn't try to analyze implicit knowledge in an S-representational way. At the relational level of interpretation, there is no question about whether or not other people are in the 'right sort' of epistemic contact with the world around them. To the extent that others can be judged as acting properly in a world-directed fashion, their attitudes are already regarded as being in touch with the world. There is only *one* way in which people can be interpreted as making contact with the world so as to succeed in making (implicit) doxastic commitments. This one way of making epistemic contact with the world is *ipso facto* the right way.

The Relational Model implies that the implicit knowledge we attribute to others when giving or understanding factive reason explanations, is a form of *common* knowledge: knowledge that expresses a shared perspective on the world, the world as we all experience it to be. In the literature on social cognition, such the implicit assumption of shared perspectives has been characterized as a form of 'projection' or 'egocentric bias' (e.g. Van Boven et al. 2000, Van Boven and Loewenstein 2003, Kawada et al. 2004, Goldman 2006). These descriptions all seem to presuppose an S-representational understanding of our tendency to approach others as sharing our view on the world. The Relational Model provides a different understanding of this phenomenon: not as a matter of projecting our own subjective worldviews onto others, but rather as an attempt to relate others to a common world. As concerns knowledge attribu-

tion, our bias towards a common understanding of things has been termed the 'curse of knowledge' (Nickerson 1999, 2001; Keysar and Bly 1995; Keysar et al. 2003; Birch and Bloom 2003, 2004, 2007). But the implicit assumption of common knowledge is not a curse; it is one of the great virtues of human social life. Without it, our discursive engagements with one another would soon run aground. Of course this assumption may sometimes conceal underlying interpersonal differences, as the experimental literature cited above reveals. But in ordinary folk psychological practice, these differences need not be taken into account in advance; they are dealt with where and when the need arises (see chapter 6).

5.4 Developing a Sense for Reasons

The implicit attribution of common knowledge reveals the world as commonly known, the 'common world', as I shall call it. The common world is the world of the community to which one belongs, as described according to the public norms of reason and proper conduct. It is the world, not as it appears to anyone in particular (at a certain place and at a certain time), but *as it ought to appear to everyone* (at that place and at that time).[84] Accordingly, relational mindreading is the act of interpreting another person's world-directed attitude in terms of the way the world ought to appear to the members of one's community (no one in particular) from that person's point of view, of interpreting him, in short, in terms of the common world, viewed from that particular standpoint.

The normativity of folk psychology has long been appreciated. The holistic nature of propositional attitude ascription has lead Davidson and Dennett, for example, to stress the fact that the attribution of mental states is necessarily constrained by considerations of coherence, rationality and consistency. In an attempt to make sense of others, as Davidson once put it, "we will try for a theory that finds him consistent, a believer of truths, and a lover of the good (all by our own lights, it goes without saying)." (1970/2001a, p. 222) Likewise, Dennett has argued that in treating others as intentional agents, we have to start from the assumption that they generally tend to believe and desire what they ought to and that their actions are such as would be rational for an agent with those beliefs and desires to perform (cf. Dennett 1987, p. 49). The general idea is that our understanding of what others have done or will do is heavily

84 See section 5.5 for a more nuanced understanding of this idea.

influenced by norm-governed judgments about what they ought (not) to have done or ought (not) to do, what it made or makes proper sense to do, given the circumstances. Needless to say, there are plenty of situations in which others defy our norm-governed expectations. But on the whole they tend to listen to our normative concerns, i.e. they tend think and act as we think they ought to.

More recently, several authors have taken these considerations one step further. McGeer (2007, see also 2001), for example, directs attention to two important questions that arise once this normative dimension of folk psychology has been laid bare: (i) How do we acquire our views about what others ought to think and how they ought to act? (ii) How does it happen that others tend to comply with our normative standards, so as to make these standards a reliable tool for everyday interpretation of each other's thoughts and actions? She answers both questions with reference to the *regulative* function of folk psychology in human social practice. Through a process of enculturation, folk psychological explications teach us how we ought to think and act under certain circumstances, while simultaneously showing us what we can expect from others under similar circumstances (cf. Hutto 2008a, p. 37). According to this line of thought, folk psychology is not merely a tool for mindreading, but also for mind*shaping* (Zawidzki 2008, forthcoming). It is because our minds are shaped in much the same way through participation in a folk psychological practice, that using our folk psychology in the interpretation of each other's behavior is often successful.

There are two points I want to extract from these considerations. The first is that the normative constraints for fast and reliable folk psychological ascriptions cannot merely be *formal*, in the sense that the attitudes attributed simply be mutually coherent and consistent, irrespective of their content. In order to see this, consider a scenario along the lines of Davidson's (e.g. 1973/2001b) 'radical' interpretation, a scenario in which an interpreter is confronted with the task of making sense of people from an alien culture, speaking a language, having customs, performing rituals, etc., that are unknown to the interpreter and vice versa. The interpreter is faced with an interpretative problem. Failing to make any sense of their actions, the only thing he can do is step up and ask them for their goals and reasons. But he doesn't speak their language and they don't speak his. In order to understand what they are saying he needs to know what their thoughts and actions are directed at, but in order to know *that*, he needs to be able to understand what they are saying. The only way for him to proceed, as Davidson pointed out, is to make certain charitable assumptions about the worldview of his interlocutors and the degree of rationality of their thoughts and actions and to interpret their utterances on the basis of these

assumptions, continuously adjusting or discarding these assumptions and adding new ones as he proceeds in the light of new behavioral evidence, until he reaches sufficient levels of interpretative success (measured by his ability to adequately explain and anticipate their actions, successfully coordinate his actions with theirs, etc.).

The protagonist of our story is faced with a genuine, practical problem of epistemic holism: in an attempt to make sense of his interlocutors, he has to start with substantial assumptions about their worldview, assumptions which are severely underdetermined by the behavioral evidence and which furthermore are in constant need of revision as they fail to provide accurate predictions of further accumulating evidence. What this shows is that, in the absence of any positive idea about *what* the minds of others are directed at, assumptions about the mere formal coherence and consistency of their thoughts, utterances and actions get us close to nowhere in the process of interpretation. The standards of rationality that guide ordinary folk psychological interpretation need to be of a *material* nature, essentially pertaining to the contents of the attitudes ascribed, telling us what counts as a proper reason for adopting which goals and performing which actions under what circumstances.[85] In this respect, the normativity of folk psychology is perhaps better characterized in terms of *canonicality* (Bruner 1990, p. 37). Characterizing folk psychology as a canon of social practice brings out the fact that folk psychology is a substantial corpus, shaping our understanding of and our responding to the world. According to Bruner, the canonical status of folk psychology lies in the fact that "it summarizes not simply how things are but (often implicitly) how they should be." (ibid. p. 39-40) The (substantial, contentful) *descriptions* of the world that folk psychological explications provide, in other words, imply *prescriptions* as to how to think and act under particular circumstances.

Fast and reliable folk psychological interpretation requires a good sense of the material adequacy of the thoughts and actions of others. The first point I wish to highlight here, is that the Relational Model actually predicts this. At the relational level of interpretation, stating how things are *implies* a commitment as to how things should be stated. There is no way of prying apart the descriptive and prescriptive elements of other people's sayings about the world or their responses to it; the idea of *mere* formal coherence and consistency, in abstraction from material adequacy, is not intelligible from this interpretative

85 See De Bruin and Strijbos (2010) and Strijbos and De Bruin (2012) for a model of folk psychological interpretation according to which reason discourse is a process of endorsing, constructing or rejecting material inferences (cf. Sellars 1953, Brandom 1994) that feature the suggested reasons for action in the antecedent and the actions to be explained in the consequens.

point of view. For the relational mindreader, the canonicality of folk psychology reflects the fact that folk psychological interpretation is constrained by our common and only appreciation of the world. Accordingly, the norms of folk psychology mirror the nature of the common world, the perceptual reports that it solicits and the inferences and ensuing actions that it licenses – the world as it ought to be perceived, conceived of and responded to.

The second point I want to extract from the above considerations regards the ontogenetic development of our capacity for goal-reason attribution. Children do not come equipped with an acute sense of material (in)adequacy of thought and action. Not even the staunchest nativist would dare to claim that children have an innate module that informs them about folk psychological norms of reason and proper conduct. What is considered proper acting or thinking is heavily dependent on socio-cultural factors. The ontogenetic question, then, is *how children learn the material rules of the game of giving and asking for reasons*.

Traditional TT and ST accounts have paid surprisingly little attention to this question. TT claims that folk psychological interpretation is guided by lawlike psychological generalizations of the kind "if A wants P and believes that doing q will bring about p, then ceteris paribus, A will q." (Borg 2007, p.6; see chapter 2.3) But such 'central action principles' get us nowhere without any positive idea about the contents of P, q and p, respectively. As we have seen, this has been considered one of the major drawbacks of TT compared to ST. Heal (e.g. 1998) took it to be an a priori truth that we identify the contents of other people's mind by means of co-cognition (see section 5.3). Gordon (e.g. 1996) and Goldman (e.g. 2006) proposed further empirical hypotheses as to how this is supposed to be accomplished: by imaginatively identifying with them through performing an 'egocentric shift', and by introspecting on pretend states about their situation, respectively (cf. chapter 4.2 and 5.2). Yet such stories are simply insufficient to explain the ontogeny of mindreading. For the nascent child *does not yet have* any substantial, inferentially articulated worldview to draw from for the purpose of co-cognition, imaginative identification or introspection. The question that needs to be answered is how children acquire a proper sense of *when to do what and why*.

To this end, consider Hutto's (2008a, 2008b, 2009) 'Narrative Practice Hypothesis (NPH). As indicated in chapter 2.3, Hutto characterizes adult folk psychology as a narrative practice, a primarily second-person practice of telling and listening to the stories behind people's actions. He terms these stories 'folk psychological narratives': stories about people (or other intentional agents) acting with goals and for reasons. According to the NPH children learn to

wield our folk psychological concepts of belief, desire and other psychological attitudes through participation in such narrative practices. Here one can think of fictional stories as found in children books, but also of non-fictional stories about actual agents. The narratives themselves, as socio-cultural objects, exemplify the core structure of our folk psychology, in the form of explications, explanations and predictions of the protagonist's actions (e.g. "Little Red Riding Hood entered her grandmother's house because she thought that the wolf was her granny and she wanted to bring her granny a basket of goods because she was ill…"). The idea is that through attentive listening and active participation (e.g. asking questions), children gain familiarity with the way psychological attitudes interrelate and lead to action. The folk psychological narratives children are exposed to can serve as exemplars for constructing stories of their own as to why someone acted in a certain way. These narratives can then again function as objects of joint attention with caregivers, who can correct and further specify them so as to meet the proper standards of folk psychology. As a result, children get bootstrapped into the folk psychological practice of telling stories about people's actions, their goals and their reasons.

What I want to suggest here, is that the NPH can function as a promising account of how children acquire a sense of the material adequacy of goals and reasons for action. Accordingly, it is through continuous practice and training in telling folk psychological narratives that they get familiarized with the specific norms of proper thinking and conduct that characterize their folk psychological community. Constant interaction with their caregivers ensures that they are slowly but surely being pulled up into the space of reasons and learn the rules of discursive engagement, of what (not) to think, say or do under which circumstances and why. The substantive worldview of their caregivers, specifying how things are or ought to be done, provides the proper 'deontic scaffolding' (Strijbos and De Bruin 2012) that enables children to bootstrap themselves into a proper discursive understanding of human social life.

Hutto, however, targets a different explanandum. His aim is to provide an ontogenetic account of our competence in belief-desire psychology. Thus, he explains:

> the stories in question serve as exemplars and teaching tools: in their guided encounters with such stories children come to see the relations that hold between the various psychological attitudes – crucially, but not exclusively, the focus is on beliefs and desires […] the way beliefs and desires conspire to motivate action – which, in abstracto, we might think of as the core folk psychologi-

cal schema, is a constant feature of these narratives. (2008a, p. 29)

But it seems the NPH could remain largely intact if we were to add a relational stratum of goal-reason psychology in the developmental timeline. The idea that children need to learn how propositional attitudes interrelate to motivate action applies just as much at the relational level. Recall Ratcliffe's point that "Just as one can say 'if B believes p and desires q, all things being equal, B ought to do r', one can say 'if p is the case and q is the case, all things being equal, B ought to do r'." (2007, p. 97; see section 4.1) It is certainly possible that children first learn the complex relation between actions, goals and reasons by getting a hold on the relational attitudes of the protagonists of FP-narratives, relational attitudes of (pretend) knowledge that link the agent to her reasons (in the counterfactual story) and relational attitudes of (pretend) desire and intention that link her to her goals. Accordingly, children's first grasp of goals and reasons would reflect the public norms of action: when one ought (not) to do what and why.

This is not only possible, but also very plausible when compared to the alternative offered by the BD-Model. In order to see this, it should again be realized that when children are first being invited to participate in discursive practice by their caregivers, they are still in the process of *learning* the material constraints of folk psychology. They do not yet have a substantial, propositionally articulated view of the world at this point in their interpretative careers; they still lack the capacity to assess the propriety of the inferential connections between thoughts, words, and actions. With this in mind, consider how needlessly confusing it would be for children at this ontogenetic stage to gauge the minds of others in a fundamentally S-representational way, to provide each and every example of proper reasoning or conduct set by their caregivers with a silent comment on the fact that this merely reflects *their* view on things, not necessarily the correct view. Such S-representational understanding would be superfluous; it could have no effect on their initial learning curve. The reason is simple: there is nothing they can contrast the set examples with, no properly worked out alternative to present or pursue in its stead. This points towards a more fundamental worry. For it could be argued that for the same reason, such S-representational thoughts *could not even make sense* to children at this stage in their social development. If there is no contrast class of possible alternatives to the goals and reasons explicated or exemplified by their caregivers, it seems the S-representational distinction between a subjective and an objective understanding of the world simply could not apply to their assessment of the situation.

The Relational Model can avoid these difficulties. At the relational level of understanding, there can be no question about the normative force of the rules of common discursive practice; the goals and reasons revealed by others can only appear as part of a given reality.[86] Children's first encounters with discursive practice are guided attempts at relating to the common world as it is pointed out to them by their caregivers.[87] Here, the activity of 'pointing out' can be taken quite literally as pointing, with words, to certain salient features in the common world.[88] The capacity to initiate and understand pointing gestures as a means for engaging in cycles of joint attention starts to emerge around infants' first birthday and is robustly present at the age of 15 months, as it begins to manifest itself in the learning and use of words (Tomasello et al. 2005). Once children begin to actively participate in conversation, roughly from 2 years of age onwards, objects of joint attention, which they have already explored in numerous non-discursive, affective, perceptual and action-oriented ways, can slowly but surely start to appear as propositionally articulated events, facts and states of affairs: the stuff that goals and reasons are made of.

86 A 'given' in the innocuous sense of the term, shaped by a linguistic practice, not a mythical, pre-conceptual 'Given' that is supposed to justify our conceptual norms (see chapter 3.2).

87 The experimental findings of Rakoczy et al. (2008; see also Rakoczy et al. 2009) appear to be a case in point. On of their experiments involved a newly invented game called 'daxing.' In the first phase of the experiment, one of the experimenters brought out some building blocks and she and the child performed some usual actions with them. In the second phase, the experimenter announced that she was going to show the child a new game involving the building blocks, called 'daxing', and explained the rules. The experimenter and the child then played the game, taking turns for a while. In the third stage, a puppet called Max appeared who had previously been introduced to the child. In the experimental condition Max announced to play the game, whereas in the control condition he announced not to play the game but rather to build. The 3-year-olds intensely monitored Max's moves and displayed distinctively normative interventions when he did not play by the rules of 'daxing (e.g. 'No it does not go like this!' or 'No, don't do it that way!'), and they did so only in the experimental condition. The 3-year-olds thus appeared to have a clear awareness of the normativity created in simple conventional rule games. For our purposes, what is interesting is that they absorbed the rules as explained by the experimenter without question or argument and afterwards tried to ensure that the puppet played in strict accordance with them. This active focus on conventional norms laid out by adults makes much sense on the ontogenetic picture sketched here.

88 Cf. Heal (2005): "Words are [...] an immensely delicate and useful way of pointing. Pointing itself is an elaborated way of focusing shared gaze. And what in turn grounds the whole enterprise is the sense of living together with another, a sense which perhaps shows itself already in the infant in those very early episodes when infant and carer smile at each other." (p. 39)

5.5 The Limits of Relational Mindreading

In section 3 I dismissed a worry regarding the inaccuracy of relational mindreading. The worry was that lack of agential similarity between different members of society, even of the same socio-cultural group, makes relational mindreading too unreliable as our dominant interpretation technique in daily social life. Differences in background, status, occupational role, character, etc. are profound. Due to these differences, the interpretation of others on the basis of one's own understanding of the world would often miss its mark. Against this, I held that there must be a vast shared background underneath all these differences, if discursive engagement with one another is ever to get off the ground. The case of Davidsonian 'radical' interpretation in section 4 further exemplified this point: A common language is non-negotiable for folk psychological interpretation.

This section serves to show that the limits of relational mindreading need not be set by the strict requirement of agential similarity. I discuss two ways in which the relational mindreader can relate to another person without relying solely on his or her own understanding of the situation: by means of assessing the situation relative the other person's (i) spatio-temporal point of view and (ii) socio-cultural characteristics.

As to (i): Consider the capacity of so-called 'level 2 visual perspective taking' (e.g. Flavell 1974). Children are said to have acquired the competence for level 1 perspective taking when they are able to understand that different people with different lines of sight may see different things (e.g. person A can see a box in front of a wall, while another person, standing on the other side of the wall, cannot see the box). Level 2 perspective taking furthermore requires that one understands that different people with different lines of sight might see one and the same thing in different ways (e.g. person A sees the box in front of the ball, while person B sees the box behind the ball). Level 2 perspective taking constitutes what Perner and colleagues (Perner et al. 2002, Perner et al. 2003, Perner et al. 2005) term a 'perspective problem'. Roughly, a social situation constitutes a perspective problem iff there are at least two intentional states involved the content of which cannot be joined by simple conjunction to yield a consistent representation. The content of person A's state 'the box is in front of the ball' cannot simply be conjoined with the content of person B's state 'the box is behind the ball', for that would result in the inconsistent proposition that the box is both in front and behind the ball. Proper understanding of the situation requires that one conjoins the contents of the intentional states of persons A and B *by making reference to their different*

perspectives: the box is behind the ball *from A's point of view* and the box is behind the ball *from B's point of view*. Level 2 perspective taking requires that one grasp the possibility of different people being intentionally directed at the same thing under a different 'mode of presentation' (cf. chapter 4.4).

Now, does relational mindreading allow for level 2 perspective taking? There seems to be no reason why it could not. Level 2 perspective taking only requires that one can assess the public world from different spatial coordinates; it does not involve the capacity to assess the public world *as privately conceived* from different spatial coordinates. Perner et al. (2002) make a distinction between 'truth compatible' and 'truth incompatible' perspective problems. Level 2 perspective taking is a truth compatible perspective problem, because the truth of the conjoined contents of the respective mental states can be preserved with reference merely to spatial coordinates of their bearers. They consider the standard 'false belief task' to constitute a truth incompatible perspective problem. In one classical experimental setup, the task involves a character (a puppet), who places an object (chocolate) in a particular location (a box) and then leaves the room. In his absence the chocolate is moved to another location (a cupboard). The subject, who has watched the scene unfold, is then asked where the puppet will look for the chocolate when it returns. In order to succeed in this task (i.e. predict that the puppet will look for the chocolate in the box) the subject must understand that the puppet still thinks that the chocolate is in the box, and therefore, it is often argued, that the puppet has a false belief about the location of the chocolate (cf. Wimmer and Perner 1983). Succeeding in standard false belief tasks constitutes a truth incompatible perspective problem, according to Perner et al., because the contents of the relevant intentional states ('the chocolate is in the cupboard', 'the chocolate is in the box') cannot be conjoined to yield a consistent representation of the scenario, not even with reference to the different points of view of the test subjects and the puppet. The proposition 'the chocolate is in the cupboard from my point of view and the chocolate is in the box from the puppet's point of view' seems inconsistent.[89] In general, the defining criterion of a truth incompatible perspective problem is that the contents of the relevant states cannot be consistently conjoined within one and the same 'possible world' or logical universe (cf. Perner et al. 2002, p. 1465) Consistently conjoining the contents of the relevant mental states demands that we make reference, not merely to points of view in one and the same world, but to points of view *in*

89 It is questionable, however, whether this is the best way to approach children's understanding of the standard false belief task (see appendix).

different possible worlds.

Perner et al.'s distinction between truth compatible and truth incompatible perspective problems provides us with a useful tool in drawing the limits of relational mindreading. In truth compatible perspective problems, the contents of different perspectives can be made compatible within one 'possible world'. For the relational mindreader, there is only one possible world that can apply to the actual world, only one frame of reference for assessing the world-directed attitudes of others: the common world of public assessment.[90] Relational mindreading can tackle perspective problems insofar as solving them only requires 'walking around' in the common world, assessing the modes in which it presents itself from different spatio-temporal points of view. In order to solve truth incompatible perspective problems, one needs to understand that agential perspectives that depict *counterfactual* scenarios can inform agents' *world-directed* attitudes. One needs to be sensitive to possibility that people's subjective worldviews may misrepresent the common world, that their private understanding of the world may be incompatible with the public rules of proper thought and action. This goes beyond the grasp of relational mindreading. It requires S-representational mindreading.

The distinction between compatible and incompatible perspectives not only applies to informational states with a mind-to-world direction of fit (cf. section 2.4). It can also be used to characterize the limits of relational mindreading in making sense of other people's motivational states, states with a world-to-mind direction of fit. The watershed between relational and S-representational mindreading is drawn by the 'satisfaction (in)compatibility', i.e. the (un)feasibility, (in)appropriateness or (un)acceptability of the motivational states attributed. As long as the desires and intentions of others do not come into conflict with our commonsense assessment of proper conduct, relational mindreading should in principle suffice.[91] In general, both the informational and motivational attitudes ascribed must be in line with our common practice of knowing, intending and acting.[92]

90 This constraint on relational mindreading is what I termed the 'disjunctive constraint' in chapter 4.4.

91 Arguably, attribution of informational states is more tightly constrained than attribution of motivational states. Even the most neutral knowledge claim seems to imply a public commitment regarding its truth, for example, whereas the performance of an insignificant action might only imply its *compatibility* with the public commitments of proper conduct.

92 Let it be noted in passing that the distinction between informational and motivational states has limited application at the level of relational mindreading. There are situations in which the specific circumstances of an agent dictate that he perform one and *only* one particular kind of action. Consider the following case. A child is drowning in the middle of a lake and there is only one person nearby who can rescue her, a man, say, fishing in a boat. For a bystander, the situation

As concerns (ii): Relational mindreading also allows for an assessment of other people's actions relative to prevalent stereotypes, relating to socio-economic class, professional role, social status, etc. Not everybody is treated alike in social practice. Immigrants, doctors, pop stars, children, scientists, prime ministers, etc., all occupy specific socio-economic positions and/or have specific social roles to play, and with each position and role come different rights and responsibilities. Scientific experts are granted knowledge of things that other people do not understand; doctors are entitled to perform procedures no other person is allowed to; children are granted much higher degrees of naivety and irresponsibility than adults, etc. These signatures of entitlements and commitments are carved into human social life, forming an integral part of our discursive practices. What counts as a proper goal or reason for action depends not only on the circumstances one is in, but also on the position or status one has and the role one is supposed to play under those circumstances. But this need not pose any problems for the relational mindreader. The world other people are interpreted as being related to, is not a projection of one's own private preferences and concerns, it is a *common* world, the world as characterized by *public* standards of reason.

This is an important point. One of the most frequently voiced objections to the Simulation Theory is that it cannot account for interpretation in cases of agential dissimilarity (e.g. Churchland 1991, Nichols and Stich 2003, Weiskopf 2005, Zawidzki 2008). To the extent in which the other person's mind is dissimilar from one's own, using one's own mind as a model for interpreting the other is deemed unreliable. But why would the resources for mindreading have

at that particular moment may very well allow one and only one proper course of action for this man: e.g. to paddle his boat towards the child in order to rescue her. A mere relational mindreader would be able to interpret this person as intentionally directed at the situation at hand only insofar as he acts in line with the way he ought to; any behavior not conducive to his paddling towards the child would be regarded as a failed action. In this case, the informational/motivational distinction is futile. There is no intelligible way for the relational mindreader to separate the informational aspects of the action (that the child is drowning and that he can rescue her by paddling towards her and getting her out of the water) from the motivational aspects (that he ought to paddle towards her). The person at the scene is interpreted, if at all, in terms of a relational 'besire' (Altham 1986, cf. McDowell 1978/1998), a state with both directions of fit with respect to different aspects of the world: knowledge-like with respect the fact that the child is drowning, etc., desire/intention-like with respect to his rescuing her by paddling his boat towards her. Prying these two aspects apart would require S-representational mindreading. It would require that the interpreter take into account "that it is at least possible for agents who are in some particular belief-like state not to be in some particular desire-like state; that the two can always be pulled apart, at least modally." (Smith 1994, p. 119) For it to be intelligible that the person in our example could intentionally respond to the child's drowning by, say, paddling away or casting a line, the bystander would have to assess the action from an essentially subjective point of view. A principled distinction between informational and motivational states can only be drawn from the perspective of someone who sees the relation between agents and the world as being mediated by their subjective representations of the world.

to be restricted to one's *own* model of the world? Why couldn't interpreters use the *common* mind as a model for making sense of others? Relational mindreading is a form of 'understanding others from the inside' (Heal 2000). The crucial question, however, is how we should interpret this phrase. From inside of *what*? Not, I suggest, from inside our own minds, as on Goldman's (e.g. 2006) introspectionist version of the Simulation Theory, nor from inside the minds of others, as on Gordon's (e.g. 1996) 'radical' version, but rather from inside the community of human beings to which one belongs (cf. McGeer 2007).[93] The common world is not the mere overlap between projected egocentric points of view; it is the world of a community of people, shaped by culture, crafted by narratives, informed by theories. Starting from our ability to co-cognize, in Heal's (1998, p. 483) sense of exercising "the same underlying multifaceted ability to deal with some subject matter" (see section 4.3), relational mindreading can draw from a rich tradition of folk psychological knowledge that specifies goals and reasons relative to socio-cultural characteristics.[94]

There doesn't seem to be an a priori constraint on the extent to which the public rules of thought and action can be tailored to fit the socio-cultural characteristics of specific (groups of) individuals. Strongly put, there is no dissimilarity between two individual human agents that could not in principle be incorporated into public life so as to render relational mindreading sufficient for making sense of each other's differences. When it comes to the limits of relational mindreading, the question is not so much *where* but *how* to draw the line. At the relational level of understanding, marking goals and reasons for specific individuals proceeds by zooming in on their socio-economic position, professional role, etc. from the *public* point of view. Whenever they act in defiance of the public norms, the relational mindreader hits the rock bottom of common ground and can dig no deeper. The line is drawn where making sense of the minds of others requires approaching them from their *private* perspective on the world, the world as represented by them in particular, irrespective of their socio-cultural characteristics, and not necessarily by anyone else.

93 Or from inside communities within the community. It is certainly possible that relational mindreading draws from different public frames of mind, depending on the social group or subculture in which our discursive engagements take place.

94 Cf. Hutto (2008a): "Stories [...] help to shape our common cultural expectations, making us familiar with the norms governing actions in "ordinary" situations. [...] Through them we learn, for example, about the social roles that pervade our everyday environments [...]." (p. 37) Heal (1995) also acknowledges a role for 'information rich' folk psychological knowledge found theories and narratives in addition to simulation as co-cognition. Thus se states that: "it is obvious that we do derive from experience, literature, political treatises, books on psychology, etc. a great deal of explicit and implicit knowledge about human nature, people's psychological states, how they arise and interact and so forth." (p. 46)

5.6 Conclusion

This chapter targeted the first part of second challenge set for the Relational Model: to show why relational mindreading is not just a conceptually coherent idea, but also a robust psychological phenomenon in human discursive practice. Section 2 revealed that the psychological plausibility of relational mindreading is not threatened by dominant explanatory theories found in the current literature. Although clearly inspired by the BD-Model, all these subpersonal accounts of mindreading allow for a relational interpretation of the explanandum. Section 3 provided two powerful considerations why such interpretation is to be preferred. As argued there, the BD-Model faces significant problems in accounting for epistemic holism and implicit knowledge attribution. I showed how the Relational Model dissolves these problems by taking a different perspective on the human discursive understanding. In section 4, acknowledgement of the holistic nature of goal-reason attribution urged us to regard folk psychology as an essentially normative practice, not only in a strictly formal sense, but also, and more importantly, in a material sense. Acquiring a proper understanding of the material (in)adequacies of thought and action is not something that children can achieve all by themselves. They need substantial input from others, people who have already mastered a subtle appreciation of the material (im)proprieties of discursive engagement and thereby have already gained a differentiated, inferentially articulated view of the world. Section 4 showed that the Relational Model either implies or gives a plausible rendering of these considerations, much more plausible, in fact, than the BD-Model. Finally, section 5 explored the limits of relational mindreading. We found that these limits need not be drawn by the boundaries of our individual preferences and concerns. Beyond agential similarity, relational mindreading can proceed as far as common sense goes.

The next chapter will approach our interpersonal differences from another direction, not by homing on stereotypical characteristics from within the public sphere, but by exploring perspectives from within the private domains of other minds. This is not as mysterious as it sounds; it is the province of good-old belief-desire psychology.

Appendix: Do False Belief Tasks Test False Belief Understanding?

In section 4 I suggested that children start their discursive careers as relational mindreaders, from around 2 years of age onwards, when they begin to actively participate in conversation and learn the rules of the game of giving and asking for reasons. This may seem at odds with recent empirical findings from developmental psychology. This appendix provides a brief survey of these findings and argues that they do not undermine the Relational Model.

Section 5 briefly discussed the standard false belief task (SFBT). Until recently, the SFBT was considered a reliable indicator of the fact that children acquire an understanding of false belief no earlier than 4 years of age (e.g., Wimmer and Perner, 1983; Baron-Cohen et al., 1985, Wellman et al. 2001). This would fit in nicely with the developmental timeline I suggested. Accordingly, children start to discursively engage with others by means of relational mindreading roughly from their second birthday onwards. When they reach the age of 4, their relational interpretation techniques are supplemented with S-representational mindreading skills, tailored to fit specific social situations, such as mimicked in the SFBT.

The SFBT relies on the subject's explicit verbal response to the experimenter's question as to where the subject will look for the object of interest. This makes it impossible to test infant's understanding of the situation before they have acquired sufficient linguistic competence to understand and answer the experimenter's questions. Recently, therefore, studies were conducted based on violation-of-expectation and anticipatory looking paradigms. These studies, it has been argued, show that false belief understanding emerges at a considerably earlier age: in 25-month-olds (Southgate et al., 2007), 15-month-olds (Onishi and Baillargeon, 2005), and even 13-month-olds (Surian et al., 2007).

Onishi and Baillargeon (2005), for example, conducted an experiment in which 15-month-old infants were familiarized with a protagonist hiding a toy in one of two locations. The protagonist left, and the toy was moved without her knowledge. Then the infants were shown scenes of the protagonist searching for the hidden toy either where she falsely believed it to be, or where it was actually located. Onishi and Baillargeon found that 15-month-old infants reliably looked longer at those scenes in which the protagonist searched at the correct location despite their false belief about where the toy was hidden, and thus expected the protagonist to search for the toy where she believed it was located. Onishi and Baillargeon concluded that this measure of action expectancy or anticipation in fact demonstrated an early understanding of false belief.

Follow-up experiments have attempted to support this conclusion. Southgate et al. (2007), for example, tested anticipatory looking in infants of 25 months old by means of a task that was quite similar to the one used by Onishi and Baillargeon (2005). Infants observed a protagonist who witnessed a puppet bear hiding a ball in one of two boxes. Then the protagonist became distracted and turned away from the scene. Meanwhile, the bear removed the ball from its original hiding location. Southgate et al. (2007) found that, on the protagonist's return, most 25-month-olds correctly anticipated her behavior and looked at the location where she falsely believed the ball to be hidden. Again, the conclusion was that this demonstrated an early understanding of false belief.

The first thing to say here is that even if these experimental findings indicate proper mastery of the concept of subjective belief, this could not by itself undermine the position defended here. It is one thing to argue that infants can ascribe S-representational propositional attitudes to others in certain highly artificial experimental settings, quite another to say that it comprises their default interpretation strategy in daily social situations. Normal adults are surely able to ascribe S-representational beliefs, but that does not imply that they do so all the time. The same would apply to the story of infants and children. Secondly, however, it is seriously doubtful whether the NVFBT can actually test for propositional attitude ascription.

To start with, it is not at all clear whether the explanatory models put forward to explain infants' performance on NVFBT are committed to an interpretation of their performance in terms of full-blown propositional attitude ascription. Baillargeon et al. (2010), for example, recently proposed a modular Theory Theory account according to which infants come equipped with an innate psychological-reasoning system that consists of two sub-systems: sub-system 1 and sub-system 2. Sub-system 1 (SS1) enables infants to register both motivational states and 'reality-congruent informational states' to other agents, and is well in place by the end of the 1st year. Reality-congruent informational states, they explain, specify what accurate information the agent possesses about the scene. Motivational states, by contrast, are defined as states that specify the agent's motivation in the scene and include goals and dispositions. Sub-system 2 (SS2) goes beyond SS1 in that it also enables infants to attribute 'reality-incongruent informational states' to another agent, and becomes operational in the second year of life (cf. Scott & Baillargeon 2009). Baillargeon et al. argue that the findings discussed above are indeed indicative for implicit false belief understanding, and they explain this in terms of SS2. At the same time, however, they merely characterize this understanding as the

ability to attribute 'reality in-congruent informational states'.

This obviously falls short of the much more advanced capacity to attribute *propositional* states to others. As Apperly and Butterfill (2009, p. 957) point out, 'in terms of content [...] no study has yet suggested that infants track beliefs involving both the features and location of an object (e.g., 'The red ball is in the cupboard') or that they track beliefs whose contents can be represented only using quantifiers (e.g., 'there is no red ball in the cupboard'); or that, in tracking beliefs, they are sensitive to modes of presentation.' On the basis of these observations, Apperly and Butterfill conclude that 'whatever [infants] represent, it is not a state with propositional content.' (*ibid.*)

Zawidzki (2011) provides a stronger and more principled argument against interpretation of infants' performance as manifesting genuine false belief understanding. He argues that, due to its experimental design, the NVFBT *cannot* test for mastery of the concept of belief. As argued in section 3, the propositional nature of beliefs is closely tied to the holistic nature of their attribution. Beliefs and other propositional attitudes only show tenuous, holistically mediated connections to behavior and external circumstances. Ascribing a belief on the basis of which someone is judged to respond to a specific environmental stimulus only makes sense if that belief can be placed against an inferentially articulated background of indefinitely many other propositional attitudes. Now how could the NVFBT test for this? A clear manifestation of full mastery of the propositional attitudes concepts would require a fair deal of linguistic comment on the observed scenario. There seems to be no other way to make sure that the subject's understanding of the behavior of the protagonist is based on an appreciation of the latter's propositionally articulated, holistically structured assessment of the situation.

Thus, the very methods of investigation of the NVFBT resist interpretation of the experimental subjects' social understanding in terms of full-blown propositional attitude ascription. Plenty of other interpretations have been provided to account for the experimental findings, moreover.[95] The SFBT fares much better, it would seem, for it relies on subjects' explicit verbal responses. Yet here too, there are alternative interpretations possible. Let it be granted,

95 Perner and Ruffman (2005) and Ruffman and Perner (2005), for example, suggest that infants might solve the NVFBT by application of behavioral rules or by drawing associations between object, protagonist and location. Apperly and Butterfill (2009) model infants' performance in terms of 'registering' rather than full-blown belief (see below). De Bruin, Strijbos and Slors (2011) offer an enactive account of the infants' understanding of the situation, according to which infants' anticipation of the protagonist's behavior is a form of tracking affordances for others. Interestingly, the concept of affordance (Gibson 1979) can be understood as implying neither a world-to-mind direction of fit, nor a mind-to-world direction of fit (cf. section 2.4).

for the sake of the argument, that when confronted with the SFBT children explicitly reason about the protagonist's (A) future behavior. We could then reconstruct their reasoning as follows: "A *saw* that the object was at location L1 but did not *see* the change of location of the object to L2; A *intends* to retrieve the object and will base his action on what he has *seen*, so he will *look for* the object at L1." All italicized intentional attitude terms can be interpreted as referring to relational, rather than S-representational states. So there is no reason why this reconstruction could not in principle be understood from a relational mindreader's point of view.

In this context it is interesting to consider Apperly & Butterfill's (2009) alternative interpretation of infants' performance on the NVFBT. On their interpretation, infants are sensitive to the agent's belief only insofar as it *registers* the object. The notion of registering, they suggest, builds on the more primitive notion of *encountering*. Encountering is defined as 'a relation between an individual, an object and a location, such that the relation obtains when the object is in the individual's field' (p. 962). A field is defined, simply, as a certain region of space around the individual. Registering is then defined as a slightly more complex psychological relation between an individual, an object and a location. An individual is said to register an object at a location when (a) she encounters the object at the location and (b) has not since encountered it somewhere else. A registering is off target when the object registered is not located where it is registered to be. The importance of the concept lies in the connections to actions: 'One can understand registration as an enabling condition for action, so that registering an object and location enables one to act on it later […] Further, registration also can be understood as determining which location an individual will direct their actions to when attempting to act on that object' (962).

If we apply the notion of registering to the above reconstruction of the SFBT, we would get: "A *registered* that the object was at location L1; A *intends* to retrieve the object and will base his action on what he has *registered*, so he will *look for* the object at L1." The point I want to make is that tracking an agent's registering of something does not require an understanding of her mental states as subjectively representing what they are directed at. Registerings too are relational states.

So what would demonstrate children's understanding of false belief? The mere presence explicit verbal response is not enough (cf. Q: "Where will A look for the object?" A: "There!" – pointing to the location). Not even all *explanations* of the prediction of the protagonist's behavior would suffice (cf. Q: "Why do you think A will look at location L1?" A: "Because that's where he last saw

it.") Only explicit comments on the protagonist's truth incompatible commitments would seem to be sufficient (e.g. "Because he mistakenly thought that the object was still at L1.").

These considerations also bear on experiments designed to test infants' and young children's understanding of subjective preferences and desires. Consider Repacholi and Gopnik's (1997) experiment, in which 18-month-olds were shown two bowls, one of yummy goldfish crackers and one containing yucky broccoli (most infants confirmed to this characterization). In the discrepant preference condition the experimenter tasted the broccoli and said 'Mmm, that's good!' When tasting crackers, she made a disgusting face and exclaimed 'Yuck, that's awful.' After this demonstration, the experimenter held out her hand and asked, 'Can you give me some?' The 18-month-olds succeed in giving the experimenter the one that she had shown a preference for, i.e. the broccoli, even though it clearly did not match with their own choice.

Successful performance on this task does not show sensitivity to the holistic inferential connections between the relevant preference and other background mental states, no more than does the NVFBT. There are alternative explanations, moreover. On Perner et al.'s proposal (Perner et al. 2005, see also Perner and Roessler 2010), for example, infants understand the situation in terms of what they term 'objective desirability' relative to specific situations. Accordingly, 'broccoli in a grown up's month' could be judged good while 'broccoli in my mouth' considered bad. This proposal can be regarded as a specific instance of what I described in section 5.5 as interpretation relative to people's social class, role or status. And this, I argued, lies within the limits of relational mindreading.

True understanding of the concept of subjective desire, Perner et al. argue, requires that one is able to ascribe *conflicting* desires, i.e. desires with contents incompatible with the contents of one's own desires. Participation in competitive games has generally been considered a good case in point. In simple competitive games involving two competitors, the desired outcomes of both competitors are mutually incompatible. Asking the child about the other competitor's desire would seem a good test for the capacity to ascribe conflicting desires. Test results show correct answers near ceiling level at ages ranging from 3 (Rakoczy et al. 2007, Rakoczy 2010) to 5 years (Moore et al. 1995), depending on the specific experimental settings. Their mutual incompatibility notwithstanding, however, one could argue that ascription of each competitor's desire to win does not necessarily require an S-representational understanding of the situation. After all, it is one of the basic rules of game playing that, in general, each competitor who participates in a game *ought to have the desire to win*. From

the point of view of common sense, therefore, *neither* one of the competitor's desires is incompatible with the public rules of proper conduct.

A better test for understanding of incompatible desires would involve the ascription of desires considered *unacceptable* by most children. Yuill et al. (1996) designed a 'wicked' desire task, in which children had to predict the emotional state (happy or sad) of a protagonist, depending on whether or not his goal of hitting someone on the head was satisfied. 5-year-olds performed fairly accurately, but 3-year-olds generally failed to provide the correct answer. Of course, even wicked desires could be interpreted in terms of objective desirability relative to social class, role or status. This possibility cannot be ruled out in advance. As in the case of testing false belief understanding, it seems that the only way to demonstrate genuine S-representational understanding of preference and desire is to invite subjects to make explicit judgments on the incompatibility between the protagonist's desires on the one hand and the common desirability characteristics of the situation on the other.

Simple non-verbal tasks such as the NVFBT or Repacholi and Gopnik's discrepant preference task cannot test for propositional attitude ascription. As far as the experimental findings on these non-verbal tasks are concerned, there is nothing that can undermine the developmental story suggested in section 4. Nor does successful performance on verbal tasks like the SFBT and mentioned conflicting desire tasks unequivocally indicate mastery of the concepts of belief and desire. This does not mean, of course, that 4 to 5-year-olds are still blind to the private dimension of other minds. What it does suggest, however, is that tracking objects for others or predicting happiness or sadness in relatively simple, constrained scenarios, is not the best way to test mastery of the concepts of belief and desire. We should start looking elsewhere for the proper application of our S-representational propositional attitude concepts.

6

The Social Functions of Belief-Desire Psychology

6.1 Introduction

According to the Belief-Desire Model of folk psychology, our commonsense understanding of each other as rational agents is essentially mediated by our mature, S-representational concepts of belief and desire. Interpreting someone as adopting goals in the light of reasons is an act of ascribing desires representing these goals and beliefs representing these reasons. On this picture, competent use of belief-desire psychology is an absolute requirement for engaging in discursive practice; without it we would not be able to regard each other as rational agents. The social function of belief-desire psychology, so it appears, is simply to enable us to discursively interact with one another.

The Relational Model gives a more nuanced understanding of the conceptual structure of our folk psychology. Accordingly, our folk psychology reveals the minds of rational, discursive beings first and foremost as public minds, minds intentionally directed at the world as characterized by established socio-cultural norms of reason. Folk psychological interpretation starts from the assumption that people think and act in line with the norms of common practice; it proceeds by relating them to the goals and reasons out in the common world that present themselves as particularly salient in the situation at

hand – goals and reasons that they ought to have, given the circumstances. On the Relational Model, therefore, mastery of the concepts of belief and desire is not an absolute requirement for proper engagement in the practice of giving and asking for reasons. This does not mean that belief-desire psychology is of no or only of limited practical significance for human social interaction. It is to say, however, that its role in folk psychological interpretation is essentially complementary.

The previous chapter revealed the importance of relational mindreading in folk psychological understanding. This chapter focuses on the many complementary functions of S-representational mindreading in human social practice. Together, this should suffice to successfully address the second challenge for the Relational Model laid down in chapter 1 and to establish the empirical validity of the distinction between a relational and an S-representational conception of the discursive mind.

Section 2 starts from the observation made in the previous chapters that S-representational mindreading allows us to interpret the actions of others even when their attitudes fail to align with the common world. The question is why the attribution of such discrepant propositional attitudes is of practical value for social interaction. I first consider the role of belief-desire psychology in providing third-person predictions and explanations of behavior and argue that it is of limited use when deployed for this purpose. Section 3 then focuses on the use of belief-desire psychology in second-person contexts. When people act out of the ordinary, coordination of our actions may become problematic and the success of joint projects may be at risk. Under such conditions it is very important that we ensure ourselves of their status as rational agents and that we are able to prevent such displays of abnormality from happening in the future. I show how subjective belief-desire psychology greatly enhances the capacity to rehabilitate rational status in the face of apparent irrational conduct and how it enables us to effectively and efficiently identify the source of other people's mistakes and to regulate their future behavior.

In section 4 I direct attention to the *critical* function of belief-desire psychology. Besides managing our discursive engagements within common practice, S-representational mindreading also allows us to discursively relate to that practice itself, from a somewhat more detached, evaluative stance. Once we start conceiving of our relation to the world in S-representational terms, common practice seizes to be our only frame of reference in judging the correctness of our world-directed attitudes. This creates new possibilities for the regulation of social practice. I briefly discuss how S-representational mindreading enables us to challenge established norms of thinking and acting and to

pursue rivaling ideas and ideals.

In the previous chapter I argued that belief-desire psychology is not our *dominant* way of making sense of each other. This chapter reveals why, in certain fairly complex social situations, it should nevertheless be considered our *best* option. Belief-desire psychology is not the conceptual basis of our discursive engagements with one another. Yet if the observations made in this chapter are correct, it does play a crucial complementary role in regulating our socially complex ways of life.

6.2 Beyond Prediction and Explanation

Relational mindreading, so I have argued, forms the psychological basis of our discursive social practice. It is the primary way through which we interpret each other's world-directedness in propositionally articulated ways. At the same time, however, relational mindreading has significant practical limitations. As explained in the previous chapters, interpretation of each other's propositional attitudes through relational mindreading is essentially confined to what is publically accessible. Contents incompatible with what one ought to say or do according to the rules of common practice cannot be regarded as the contents of world-directed attitudes – merely as the contents of suppositional attitudes. Interpreting other people's goals by means of relational mindreading only succeeds insofar as their goals are realistic and acceptable. When an agent adopts an unrealistic or unacceptable goal, the relational mindreader will soon run out of options. The relevant intention cannot be interpreted as such, for its content is incompatible with the interpreter's assessment of the ways the world could or should be changed as a result of the action. Chances are high that she will not be able to find anything in the common world to relate the agent to so as to make the action even remotely intelligible. Similar considerations apply to the attribution of reasons. Relational interpretation of other people's reasons only works to the extent that their reasons are constituted by worldly offerings that make their actions appear as acceptable, appropriate or required. From a relational point of view, the idea of conflicting views about proper reasons for action is not intelligible. False beliefs about the situation responded to or the appropriate means to achieve a goal may moreover render the action unjustifiable. When agents are misinformed in this way, it may simply be impossible to reconstruct the action as a proper response to the common world.

It has been suggested many times in the previous chapters that the virtue

of S-representational mindreading lies in the fact that it enables us to overcome these limitations. The most straightforward way to understand this is that it allows us to conceive of people's goals as the objects of their desires, and of their reasons as the objects of their beliefs. Subjective propositional attitudes are not constrained by the dictates of public evaluation. Desires may be unacceptable or unrealistic, beliefs false or otherwise inappropriate, and as such may inform people's world-directed behavior. According to this line of thought, belief-desire psychology is simply the subjective counterpart of relational goal-reason psychology. As competent belief-desire psychologists we are able to make sense of other people's actions even when their goals and reasons defy the rules of common practice. This greatly enhances our skills in making sense of others when their actions seem strange or out of line. What concerns us here is the *practical significance* of this fact. Belief-desire competence increases our ability to rationalize apparently counter-normative behavior. *But why is this a good thing?* How does it facilitate human social interaction?

The first option I want to consider is that belief-desire psychology serves third-person prediction and explanation of behavior. As indicated in chapter 1, Prediction and explanation is often considered to be the primary business of folk psychology by standard Theory Theory and Simulation Theory accounts. Davies and Stone (1995b) explicitly link the activity of generating folk psychological explanations and predictions to the rationalization of behavior. In an introduction to the TT/ST debate, they argue that

> one impressive fact about human beings is that […] they develop the capacity to deploy psychological concepts such as belief and desire in the predictions and explanations of the actions and mental states of other members of the species. These predictions and explanations are said to *rationalize* the subjects actions or mental states; they present the subject's beliefs and desires as providing him or her with reasons for acting and thinking in certain ways. (p.2, emphasis in original)[96]

In the present context, the idea would be something like this. When people act out of the ordinary, we cannot rely on our normal expectations about their future behavior. In order to keep coordinating our actions successfully, we need

96 In the current literature, this is still considered to be the consensus view shared by TT and ST proponents: "Though Theory Theorists and Simulation Theorists disagree over the process of mindreading, they *agree* that how we understand and interact with others in social environments is by explaining and predicting their behavior on the basis of mental state attributions." (Spaulding 2010, pp. 121-122, emphasis in original)

to understand why they behaved abnormally so that we can figure out what to expect next. Rationalization of the anomalous behavior of other people enables us to generate explanations that serve to make accurate predictions about their future conduct. There is certainly some truth to this, but I think that the value of belief-desire predictions and explanations for quotidian use is easily overestimated.

When an agent acts in defiance with the public norms of action, the predictive and explanatory powers of the relational mindreader soon give out. When the agent acts on a false belief, for example, the relational mindreader will likely interpret it as a mere failed attempt to perform a world-directed action. This gives him some room to speculate about the agent's future behavior: he may predict that the agent will suffer from the same malfunction in similar situations in the future. Appeal to the relevant malfunction also has some explanatory potential: at a minimum it contrasts failed world-directed actions with mere as-if actions (see chapter 4). But this is about as far as it goes. The relational mindreader's predictions and explanations are restricted to the particular type of behavior exhibited. The problem is that she cannot interpret the agent's malperformance *as informed by a propositional state*. She will not be able to speculate about the agent's thoughts and actions based on his false belief. Suppose John takes out the garbage on Tuesday, believing it is Wednesday. His Jonesian neighbor will regard this as a failed action of taking out the garbage on Wednesday. Solely on the basis of this malperformance, she will not be able to appreciate the possibility that today (i.e. Tuesday) John is likely to engage in other activities that he is used to on Wednesdays. The only way for his Jonesian neighbor to predict that, say, John will not go to work today (because Wednesday is his day off), is by association based on prior clustering of failed actions of these kinds on Tuesdays in the past. Since she doesn't understand John's action as resulting from a false belief, she is unable to judge which sayings and doings John is entitled and committed to, *relative to that false belief*. She is unable to predict that, *ceteris paribus*, John *will* do what he *ought* to do and will *not* do what he ought *not* to do, relative to his false belief that it is Wednesday today. The same goes for explanation: different types of abnormal behavior can be explained with reference to the false belief attributed in virtue of the inferential connections with other beliefs, desires and intentions. On the assumption that John is by and large a rational agent, it can be inferred that his not going to work today is probably also informed by his false belief that it is Wednesday today. For the attributer of false belief, there may be a *rational pattern* in John's abnormal behavior.

S-representational mindreading thus greatly expands our capacity to

speculate about other people's reasons when their actions are informed by false or differing beliefs. A similar story can be told about actions informed by inappropriate or unrealistic desires. Yet are the explanations and predictions we are able to generate through S-representational mindreading generally accurate enough to serve as the basis for successful interaction?

Recall the holism problem from chapter 5.3. As explained there, epistemic holism is a potential problem because it can lead to underdetermination. The interpreter cannot decide on the correct explanation of the agent's action because it is compatible with a wide range of equally (im)plausible interpretations. Furthermore, each interpretation allows for many equally (im)plausible predictions about the agent's future actions. It is a problem that becomes urgent in situations of Davidsonian 'radical' interpretation, when the interpreter cannot determine the contents of other minds with reference to the common world. In cases of radical interpretation, S-representational mindreading would be our *only* way to make sense of others, yet at the same time, it would be a very *poor* way. Rampant underdetermination would make efficient and effective social interaction practically impossible. The point of introducing the case of radical interpretation was to make clear that *we* are in no such predicament in our day-to-day social affairs. It is because we have a practice in common that we may interpret each other's thoughts and actions according to the do's and don'ts of that practice. Holism is not a problem as long as the minds of others are constrained by the common world.

We started from the observation that a subjective shift towards S-representational mindreading may help us to rationalize other people's actions when relational mindreading in terms of the common world runs out of resources. S-representational mindreading becomes practically relevant *precisely* in those situations in which the thoughts and actions of others *fail* to meet our normative expectations. Under these circumstances, however, the holism problem looms again. Perplexed by the behavior of others, we may find ourselves simply unable to come up with any rationalizing explanation, let alone to predict their next move. In less bizarre cases, the abnormal actions of others will often generate too many candidate explanations, which will make us feel hesitant to rely on any one of them in particular. John takes out his garbage on Tuesday, but it will not be taken away until Wednesday. Does John believe that it is Wednesday today? Does he intend to go away for a few days? Perhaps he wants to clean out his shed? Or is he just trying to annoy his neighbor?

One may wonder how urgent the problem of underdetermination is in the case described. Why should his neighbor *care* why John takes his garbage out a

day early? We often shrug our shoulders at the display of abnormal behavior by others. And even if we take interest in interpreting others when they act in strange or inappropriate ways, we are often more concerned with the impression they make than with providing accurate explanations and predictions. In such cases, underdetermination may actually suit our purposes rather well (see section 3). There is bound to be an interpretation of someone's eccentric or unusual behavior that confirms our prior determined evaluations (consider gossip, for example). Yet we often feel reluctant to rely on such evaluations when *predicting* their behavior is practically important. The moral is that when underdetermination is not deemed problematic, it's probably not explanation and prediction we are primarily concerned with.

This being said, there are plenty of situations in which we do feel it is important that we find out about the reasons and goals that explain the anomalous behavior of others. Under such circumstances, underdetermination is an urgent problem. As an example, consider the following case:

> As I was preparing for a long visit to St. Louis, I asked my wife to arrange for my car to be serviced and kept in my local garage while I was abroad. I supplied her with its telephone number and she kindly made the booking for me. On the morning of my flight, she agreed to drive me to Heathrow after I dropped my car in at the garage. We set off, each in our own car with her in the lead, her boot laden with my luggage. As we came up to the turning for the garage, she stopped at a set of red traffic lights but unusually failed to signal. This surprised me because my wife is a stickler for such things. But then something even odder happened. To my amazement when the lights changed she did not turn, but began driving toward the town centre, straight past the garage at which she herself had made the booking! As time was against us, this alarmed me. I raced to make sense of her action, assuming that, very uncharacteristically her mind must have been elsewhere. At first, I flashed my lights with my signal light blinking, expecting her to realise that I was no longer following and hoping that she would notice her mistake. Things went from bad to worse when I saw her cast a glance in her rear view mirror without stopping. At this point, I was faced with a rather tricky interpretative problem. Given that my wife is very competent and reliable, lacking any malicious streak or any reason to act so, I was at an utter loss to make sense of her actions. Although a number of possible explanations sprang to mind, knowing my wife, none of these looked plausible.

I was unable to make any sense of her actions. (Hutto 2004, p. 569)

What this example shows is that when shrugging our shoulders at the abnormal behavior of others is *not* an option, we tend to be very much concerned with getting our explanations *exactly right*. Typically there is something at stake; giving the wrong explanation will have important consequences. Under such circumstances, even the slightest degree of underdetermination may pose a significant obstacle for successful interpersonal coordination. The problem is that precisely in those cases in which third-person explanation and prediction of behavior in terms of subjective representational states is most needed for successful interaction, i.e. when people act in apparently counter-normative or otherwise puzzling ways, it is bound to be *least* effective. As Hutto observes: "given that proper reason explanations require us to designate the reason for acting – as opposed to simply offering a possible reason for acting – […] third-personal approaches are of limited use." (*ibid.*, p. 570)[97]

As competent belief-desire psychologists we are able to rationalize the anomalous behavior of others in ways that are systematically precluded from the mere relational mindreader. But that doesn't help us much when all we can do is speculate about it from a third-person stance. The holism problem often makes explaining and predicting the behavior of others in terms of discrepant beliefs and desires too unreliable for practical purposes.

6.3 Managing Discursive Engagements

Our ability to make sense of the anomalous behavior of other people is largely dependent upon the information they or others in their vicinity can provide. A generally efficient and reliable method for coming to grips with the unusual, strange or inappropriate behavior of others is to let them make themselves understood in conversation. Rather than trying to explain or predict their actions for them, we can ask them to explain or 'predict' their actions for us.[98] It goes without saying that this is not always possible, appropriate or even desirable. And sometimes the answers we get 'from the horse's mouth' are uninforma-

97 Here is the denouement of Hutto's story: "Luckily, this is what explained the otherwise disturbing behaviour of my wife on the way to the airport. After the incident, she explained that although it was true that she had phoned the garage to make the appointment herself, and she had used the number I had given her, she believed it was the number for our old garage, in the next village." (2004, p. 570).

98 In chapter 2, footnote 15, I indicate the sense in first-person avowals of intention can be regarded as cases of prediction of future behavior.

tive, nonsensical, evasive or (deliberately) misleading. Yet there may be other people around who can provide some of the information we need, people who know them well, perhaps better than they think they know themselves, people who may actually have spoken to them about the particular course of action that we are trying to understand. In second-person contexts, we do not have to speculate about the reasons and goals of others. When we don't understand or won't accept the answers given in response to our questions, we can ask for further clarification or suggest a different interpretation of what happened. This will normally evoke further responses by the other party, answers we may accept or again challenge, in which case additional information may be required, etc.[99] There is no guarantee that such conversations will have a satisfactory result. And sometimes we keep suspicious of the true motives of others despite the answers provided. Still, this poses no threat to the claim that in general, engaging in the game of giving and asking for reasons is by far the most efficient and reliable means of getting the information we want about others when we are unable to make sense of them.[100]

The previous section started from the observation that belief-desire competence increases our ability to rationalize anomalous behavior. The question was how this might actually benefit the interaction between people. Consider Bruner's (1990) claim that

> when you encounter an exception to the ordinary, and ask somebody what is happening, the person you ask will virtually always tell a story that contains *reasons* […] All such stories seem to be designed to give the exceptional behavior meaning in a manner that implicates both an intentional state in the protagonist (a belief or desire) and some canonical element in the culture […] *The function of the story is to find an intentional state that mitigates or at least makes comprehensible a deviation from a canonical cultural pattern.* (pp. 49-50, emphasis in original)

99 See De Bruin and Strijbos (2010) for an account that models everyday reason discourse as a process of what Brandom (1994, 2000) terms 'deontic scorekeeping': of assessing the correctness of material inferences expressed in answers to questions why against the background of the utterer's commitments and entitlements.

100 Of course, once we have been informed about the goals and reasons of others, we can use this information in order to generate further third-person explanations and predictions. Yet the information provided may not suffice for our explanations and predictions to reach the level of accuracy we feel is necessary for our interactive purposes, and so second-person adjustment may still be called for. The point is that our ability to generate accurate explanations and predictions of anomalous behavior depends heavily on information provided in conversation. Furthermore, the explanations and predictions we come up with are defeasible and readily revised in light of information gained in conversation.

Following Bruner on this point, Hutto (2008a) argues that "folk psychological narratives can function as 'normalizing' explanations, allowing us to cope with 'unusual' or 'eccentric' actions, by putting them into contexts that make them intelligible, where possible." (p. 7) In second-person contexts, explanation of action in terms of the agent's goals and reasons serves to provide a context in which the action can start to appear as a rational response. We ask the agent to reconstruct his action so as to make it appear 'normal' again, i.e. as falling within the scope of the norms of reason. In this way, folk psychology can serve as a kind of social glue. By creating a situation in which someone can make himself understood, we offer ourselves the opportunity to 'level' with him again, as one rational agent towards another. The practice of giving and asking for reasons thereby serves as a tool for re-establishing rational engagement. It helps us to create, maintain, restore, and even intensify our interpersonal relations (cf. Andrews 2007, McGeer 2007).

It often occurs that people explain their seemingly abnormal behavior by drawing attention to some feature that is easily accessible through relational mindreading. The interpreter may for example be unaware of the reason the agent responded to. An answer to a question why may explain the action simply by relating it to certain features of the common world that constitute a proper reason for performing the action under consideration. John puts on his coat, grabs his umbrella and walks towards the door. His roommate is surprised. Last time she looked out the window, 10 minutes ago, the sun was shining. She asks him why he's taking his umbrella. He replies that it is raining. She looks over her shoulder and sees it's raining now. In this scenario, the roommate can make sense of John's action of taking his umbrella by relating him to a reason in the common world: that it is raining now. Relational mindreading suffices. Problems arise as soon as the actions of others are informed by attitudes that fail to align with the common world. Answers given to questions why give expression to these attitudes and cannot be interpret as world-directed responses by the mere relational mindreader. The answers the agent provides only make him appear more out of touch with reality. In an attempt to make sense of the agent's action the relational mindreader gets more inexplicable behavior in return.

Competent use of belief-desire psychology enables us to accept a wider range of 'normalizing' explanations provided in response to each other's questions why and thereby greatly increases our options when trying to re-establish a common ground for future interaction. As S-representational mindreaders, we can ascribe propositional attitudes that vary considerably from the dictates of common practice. Yet, due to holistic nature of interpretation, this may ac-

tually serve to bring people *closer* within range of the normal again. Ascription of a small set of discrepant attitudes can minimize the overall deviation from the norms if the abnormal behavior is interpreted in a context adjusted to that set. In second-person contexts, we moreover do not have to speculate about the relevant attitudes: typically, they are expressed in the agent's answers to our questions.

In complex societies, most actions require interpersonal coordination. In order to successfully go about our business, there is a lot we need to know and a lot that needs to be done that we simply cannot find out or do all by ourselves. We heavily rely on others as extended pairs of eyes, ears and hands. In the daily pursuit of our goals, we have no other option than to act on the general assumption that others will behave according to the rules, speak the truth, keep their promises and meet their obligations (cf. Morton 2003, McGeer 2007). When others act in strange or seemingly inappropriate ways, coordination of our actions may become problematic and the success of joint projects may be at risk. Under such conditions it is often crucial that we re-ensure ourselves of their normative status as rational, reliable and generally cooperative agents, and that we are able to prevent such displays of abnormality from happening in the future. Finding a rational pattern in the seemingly counter-normative behavior of others may be crucial in order to rehabilitate their status as rule-abiding participants and to restore the cooperative spirit necessary to maintain our complex ways of life.

Rationalization moreover plays an important role in determining whether or which sanctions are appropriate for compensation and rehabilitation. Seemingly inappropriate behavior is normally assumed sanctionable, unless excusing or extenuating circumstances can be presented. If the agent's story convinces the audience, sanctions may be attenuated or avoided. Such interpretations are contestable and negotiable, however. They moreover generate new expectations. Failure to meet these expectations in the future will again damage one's status as bona fide conversation or cooperation partner (cf. Zawidzki forthcoming, ch. 7). In general, each move in the game of giving and asking for reasons is assessed against the background of commitments already undertaken or explicitly acknowledged. The rationalizations provided are subject to the public rules of proper discursive conduct and this narrows down the range of acceptable answers considerably. Still, competent use of belief-desire psychology greatly improves the options available for both parties to reach a satisfactory result in the give-and-take of reason discourse. In a Jonesian society of mere relational mindreaders, the inability to accept counterfactual stories as justificatory or exculpatory for apparent counter-normative behavior

reduces the chances of restoring the social balance. Sanctions will often be disproportionate and ineffective and full rehabilitation may remain out of reach.

The sanctioning of abnormal behavior takes place against the background of the important regulative role of folk psychology (see chapter 5.2). Throughout our lives, we correct and are corrected by others so as to think and act in conformity with public standards. From this perspective, 'normalization' does not only serve to make the abnormal behavior of others *appear* as normal, but also to *make* it normal and to prevent such displays of abnormality from happening in the future. S-representational mindreading can also play an important role in educative practices.

Consider the following case. A teacher is perplexed by the poor performance of one of her students on last week's exam. She asks him to explain to her how he reached his answer to one of the exam questions. In response to her question, he expresses a false belief that makes his approach appear even more puzzling. Being a skilled S-representational mindreader, however, the teacher is able to interpret his utterance as giving expression to a yet inexplicable belief. And this motivates her to ask further questions ('Why do you think so?' 'Please explain to me how you came to this conclusion.') Persistent as she is, the teacher soon finds the underlying beliefs that rationalize his malperformance and manages to teach the student how to solve the problem.

Now consider the predicaments of a Jonesian teacher. Since she cannot understand the student's answer as giving expression to a false belief, the whole point of asking for further clarification *of his answer* escapes her. Faced with the apparent abnormality of the student's answer, she can only try to correct him by showing him how it is properly done ('No, that's not right! You ought to do it this way!'). But the student will not understand her intervention; it doesn't make sense from his perspective. She may succeed in teaching him a new rule for giving the correct answer, but the rule will only apply in the context of the specific exam question under discussion. Below the surface of the conversation, the core attitudes responsible for his malperformance remain unchallenged. They will keep influencing his way of thinking and cause him to give incorrect answers when the question is framed differently. His teacher may again try to correct him, but each time the effect will be limited to the specific conditions set by the exam questions. Malperformances and their corrections multiply without any hope of actually teaching the student anything.

For the S-representational mindreader, dialogue is an efficient and effective means of getting at the source of other people's abnormal conduct. Sometimes the identification of a discrepant attitude expressed in the agent's response suffices for effective intervention. We can point out the error to the

agent and thereby re-establish a shared context for successful interaction in the future. But it may be necessary to ask further questions. Perhaps the attitude expressed in the agent's response is based on other discrepant attitudes, attitudes that would have remained unchallenged if we hadn't asked further questions and which might have resulted in other, seemingly unrelated forms of abnormal behavior in the future. In this way, S-representational mindreading enables us to prevent a whole array of abnormal actions through a single round of discourse.

In the preceding paragraphs I focused on situations in which the norms against which to assess aberrant behavior are relatively clear in advance. In error cases, for example, the agent normalizes his behavior by revealing his error and by showing that, relative to his error, he actually endorses the pre-established public norms. In educative contexts, students are expected to grant the teacher the authority to establish the criteria for successful performance. Here too, there is normally no disagreement about which norms are relevant in the particular context and how to apply them. Sometimes, however, re-establishing rational engagement demands "that we extend the range of what we think as falling within the scope of the 'normal'." (Hutto 2004, p. 560) Consider cases of *dispute* about how things are or what ought to be done. In such cases, it may not be obvious who is to be 'normalized' according to what norms. Any or each interlocutor may be mistaken and there is often no agreement about how this can be settled.

On a mere relational conception of mind, interpersonal conflict can only be interpreted as a matter of disobedience of one (some) of the parties involved. Disobedience calls for proper sanctions, not for discussion. As S-representational mindreaders, however, we are able to compare and contrast the perspectives of different individuals. When drawn into an argument, we can interpret the situation as a clash between conflicting subjective points of view. This helps to appreciate the possibility that *both* (all) parties may have to give in on certain points and it motives us to reach consensus, e.g. by articulating new interpretations of the norms under discussion or by restricting their applicability.

In this process of negotiating between conflicting perspectives, practical considerations largely determine whether and to what extent reaching consensus is appropriate or required. Perhaps agreement on a particular issue is necessary to guarantee the success of our cooperative efforts or to avoid certain dangers. If the issue is of no real importance, however, continuing the discussion may not be worth the trouble. And sometimes it seems better to tolerate divergence of different subjective views; insistence on convergence may lead

to awkward, hostile or even violent situations. Belief-desire psychology greatly enhances our conflict management skills. To the extent that it is deemed appropriate or required that we resolve our differences, S-representational mindreading increases our options in seeking to convince or persuade each other, by approaching matters from another direction, making further inquiries, doing further research, seeking advice from experts, etc. When it is undesirable or inappropriate to start or continue a discussion, or when the costs of reaching agreement are too high, ascription of subjective propositional attitudes enables us to explicitly mark our cognitive or conative differences so as to avoid further confrontation and to prevent misunderstanding in the future. In this way, the ascription of subjective propositional attitudes allows us to draw and redraw the delicate line between each other's public discursive responsibilities on the one hand, and our right to privately believe and desire on the other.

Let me also draw attention to the role of belief-desire psychology in what Malle et al. (2007) call 'impression management', i.e. "attempts to influence an audience's impression of either oneself or another person." (p. 495) Sometimes our interpretative efforts are not primarily directed at genuinely finding out what others really intended to do and why. The explanations we provide of other people's behavior may rather serve to confirm our prejudices and prior determined evaluations. Finding support for the stories we have told about others, for example, may be important to avoid losing face in front of our peers and to maintain our position in the social hierarchy. When people act in seemingly counter-normative ways, their behavior moreover tends to evoke all kinds of emotions, reactions and responses from us: 'reactive attitudes', as Strawson (1962) called them, of anger, resentment, disapprobation, but also of gratitude, admiration, commendation, etc. Sometimes our reactive attitudes seem out of order, to the person to whom they are directed, or to some third observing party. When this is the case, it is not only the *other* person's behavior that requires an explanation, but also the behavior *we* displayed in response. Explanations of the other person's behavior may then function as a justification of the ways in which we reacted to that behavior. Our status may be at stake, as a person of good will, a forgiving friend, a figure of authority, etc.

S-representational mindreading can be drawn upon whenever we feel inclined to display people in a more or less favorable light.[101] In all such cases, as

101 There is evidence from social psychology which suggests that spontaneously generated behavior explanations in terms of propositional attitudes are influenced by such normative considerations. Malle et al. (2007) found that adult observers are more likely to rationalize behavior in terms of propositional attitudes (rather than providing 'causal history explanations' that cite factors that did not figure in the actor's decision making) when they are motivated to portray

Zawidzki (forthcoming, ch. 7) observes, underdetermination due to holism is feature rather than a bug. Competent use of belief-desire psychology increases our options when trying to justify our prejudices and spontaneous reactions towards the behavior of others, by reconstructing it in ways that make our judgments and reactions seem more acceptable. It gives us more wiggle room to excuse the apparent misconduct of the ones we like or to condemn the seemingly innocent behavior of the ones we don't. At the extreme, belief-desire psychology can be exploited to intentionally damage the reputation of others when it is deemed conducive to reaching our goals.

A focus on second-person contexts of interpretation reveals that the epistemic task of finding out about each other's goals and reasons always takes place in a social setting in which established ways of interaction are potentially at stake. In this section I have explored some of the ways in which S-representational mindreading enhances our ability to successfully manage our dealings with one another in discursive practice. Folk psychological explanation often takes the form of mitigation, justification, exculpation, accusation, approbation, etc.[102] Competent use of belief-desire psychology gives us more leeway to find the explanation that best suits our social needs.

6.4 Evaluating Common Practice

The previous section showed how competent use of S-representational belief-desire psychology increases our options for adjusting, restoring or regulating our discursive engagements with one another when our thoughts and actions appear to go against the norms of common practice. But it seems that we can also engage in S-representational mindreading from a more detached, critical perspective: not so much in the course of managing our discursive engagements *within* common practice, but rather in an attempt to discursively relate *to* that practice itself.

It is only a short step from interpreting each other as subjectively repre-

the actor in a favorable light. Importantly, this effect persisted independently of the observer's knowledge about the actor. Consider also the so-called the 'Knobe effect' (e.g. Knobe 2003, 2006). In this case, the attribution of intentions seems to be influenced by negative normative judgments. Based on these and other findings, Pettit and Knobe (2009) argue that the impact of normative considerations on folk psychological attributions is pervasive.

102 Cf. Malle et al. (2007, p. 504): "behavior explanations serve more than an epistemic function: they are a social activity to manage ongoing interactions [...] Explanations can be used to clarify, justify, defend, attack, or flatter; they serve as tools to guide and influence one's audience's impressions, reactions, and actions."

senting the common world, to interpreting the common world, or at least certain parts of it, as the contents of our shared subjective representations. When we adopt a relatively disengaged, evaluative stance towards our everyday lives and start conceiving of common practice as being informed and maintained by our subjective attitudes, the question presents itself as to whether our common views and ways of doing things are (still) justified or whether they are in need of revision. From this perspective, common practice starts loosing its appeal as a given[103], as providing the unquestionable and incorrigible standards of truth and proper conduct. Claims with the impact of common knowledge may start to appear as expressions of shared subjective beliefs, rules with the power of moral imperatives as articulations of shared subjective values and desires. As such, questions can be raised regarding their accuracy and appropriateness. S-representational mindreading enables us to make a distinction between the public reality of common practice and its private appearance to individuals. But with this distinction also comes the contrast between public appearance and *objective* reality.

At the relational level of interpretation, public appearance *is* objective reality, or rather, objective reality has only one appearance: that of the common world. Intersubjectivity is conceived as having what Brandom (1994) calls an asymmetric *I-we* structure. On this conception, inquiry into how things are objectively presupposes "the existence of a privileged perspective – that of the 'we', or community. The objective correctness of claims and of the application of concepts is identified with what is endorsed by that privileged point of view." (p. 599) At the relational level of understanding, there is only one legitimate perspective when it comes to the objective correctness of conceptual, epistemic and practical norms, a perspective which is therefore automatically privileged. Accordingly, there is no "room for the possibility of error regarding that privileged perspective; what the community *takes* to be correct *is* correct." (ibid.) Answers to questions how things are objectively are prescribed by the dictates of the community to which one belongs; there is no way of going beyond this orthodoxy. The perspectives of individual members ought to be compatible with this perspective, on pain of not being recognized as claims to objectivity at all. For the relational mindreader, there is no conception of objectivity that goes beyond the confines of the common world, no understanding of objective truth that exceeds the intersubjective truth of common knowledge.[104]

103 A given, not a Given; see chapter 5.4, footnote 86.

104 In this context, recall Davidson's claim, quoted in chapter 5.3, that "Communication

Once we draw the distinction between subjectivity and objectivity at the level of the group or community as a whole, however, the identification of objective correctness with the norms of common practice can no longer be maintained. The subjective nature of the public view implies the possibility and legitimacy of competing views. The common world is revealed as one particular, and possibly mistaken, conception of how things really are. From an S-representational perspective, intersubjectivity can be perceived as having a symmetric *I-Thou* structure (Brandom 1994, p. 599). On this conception, each perspective is at most *locally* privileged as to how things really are or ought to be done. The 'we' of common practice starts to appear as another 'you', whose views are disputable in principle. From this critical stance, any *global* claim to objectivity looses credibility. The symmetry of the *I-Thou* distinction

> ensures that no one perspective is privileged in advance over any other. Sorting out who should be counted as correct, whose claims and applications of concepts should be treated as authoritative, is a messy retail business of assessing comparative authority of competing evidential and inferential claims. Such authority as precipitates out of this process derives from what various interlocutors say rather than from who says it; no perspective is authoritative as such. There is only the actual practice of sorting out who has the better reason in particular cases. (ibid., p. 601)[105]

depends on each communicator having, and correctly thinking that the other has, the concept of a shared world, an intersubjective world." (1982/2001c, p. 105). Davidson fails to draw the distinction between this concept of an intersubjective or common world and the concept of objective reality, however. For the text continues: "But the concept of an intersubjective world is the concept of an objective world, a world about which each communicator can have beliefs." (ibid.) Elsewhere Davidson states that "Thought, propositional thought, is objective in the sense that it has a content which is true or false independent (with rare exceptions) of the existence of the thought or the thinker." (1997/2001c, p. 129) This statement can be read in two ways. On the first reading, it says that thought is objective insofar as its content is true or false independent of any particular instance of thinking. This amounts to the idea that thought has objective purport in the sense of being directed at or about something in the external world. On the second reading, it says that thought is objective in the sense that it has a content which is true or false independent of the *community of thinkers* to which the thinker of the thought belongs. The first reading is implied by the concept of the common world, the second, however, is not. The common world implies the existence of a community of rational agents who inhabit it. The idea of the common view being false, of the world objectively being a certain way without *anyone* in one's community thinking or being disposed to think about it in that way, is not intelligible from a relational point of view. This notion of objective reality can only be grasped from an S-representational stance.

105 Brandom proposes to construe objectivity "as consisting in a kind of perspectival *form*, rather than in a nonperspectival or cross-perspectival *content*. What is shared by all discursive perspectives is *that* there is a difference between what is objectively correct in the way of concept application and what is merely taken to be so, not what it is – the structure, not the content." (ibid. p. 600) Notice that this formal understanding of objectivity is only intelligible from an S-representational point of view. At the relational level of interpretation, objectivity must have

If the Relational Model is correct, this thoroughly S-representational conception of intersubjectivity does not form the interpretative basis of human social practice. In most of our day-to-day interactions, the common world suffices as the referent of and standard for our thoughts and actions. Adopting an S-representational *I-Thou* stance towards our rational engagements with one another requires considerable training and effort. It is something we can do only some of time and each time only with regard to certain aspects of our lives. Yet it plays a crucial role at the more advanced levels of the game of giving and asking for reasons. A critical attitude towards the epistemic norms of common practice is what drives scientific inquiry, for example. And questioning the presumed unfeasibility or inappropriateness of certain aspirations can mark the beginning of technological development or of social and political reform. Closer to home, critical assessment of inherited values may be necessary to resolve personal conflicts or to make suitable changes in lifestyle. And overcoming interpersonal differences often demands that we re-evaluate the norms that we implicitly endorse. In these and numerous other ways, competent use of belief-desire psychology enables us to go beyond the practice of commonsense and adopt a critical attitude towards the relation between mind and world.

6.5 Conclusion

According to standard accounts of folk psychology, belief-desire psychology is the bread and butter of human discursive practice. Without it, no one could treat any other as adopting goals and doing things for reasons. On the Relational Model presented here, belief-desire psychology is more like a precision instrument in our folk psychological toolkit. It is not used for our mundane dealings with one another, not even when these interactions take place at the discursive level of giving and asking for goals and reasons. Our mature, S-representational concepts of belief, desire and other propositional attitudes play an essentially complementary role in commonsense social understanding; they are specifically designed to deal with non-standard, unexpected, difficult or otherwise problematic social situations.

In this chapter I explored some of the ways in which human interaction can benefit from the subjective shift towards belief-desire psycho-

substantial content: the world is given objectively as it is publically conceived.

logy. S-representational mindreading increases our capacity to make sense of seemingly irrational, counter-normative behavior. By ascribing a limited set of discrepant attitudes, the behavior displayed can be made to fit a rational pattern so that the overall deviation from the norms of common practice is reduced. I argued that this is of limited use for strictly third-person prediction and explanation. In second-person contexts of interpretation, however, it greatly facilitates the management of our discursive engagements with one another. It increases our capacity to ask, give and understand 'normalizing' explanations that seek to maintain or rehabilitate people's normative status or to determine proper sanctions for norm violation. It plays a crucial role in the efficient and effective correction of mistakes and the regulation of future behavior in educative contexts. Competent use of belief-desire psychology also serves our conflict and impression management skills. Approaching dispute from an S-representational point of view greatly increases the chances of finding a compromise that is acceptable for all parties involved. And the ascription of discrepant attitudes gives us more interpretative options when reconstructing behavior so as to fit with the impression others make on us or to alter the impression we make on others. Finally, I showed how a subjective conception of mind influences our thinking about objectivity. From an S-representational point of view, the givenness of common sense may start to appear as the mere public appearance of an objective reality it is supposed to be about. S-representational mindreading thereby enables us to adopt a critical stance toward common practice and to evaluate, challenge and even change the conceptual, epistemic and practical norms that shape our ordinary ways of life.

In the next and final chapter, I will show how these considerations help to explain the reflective fallacy introduced in chapter 1: the fallacy of projecting certain philosophical analyses of intentional action in terms of beliefs and desires onto our spontaneous, commonsense understanding of each other as intentional agents who perform goal-directed actions for reasons.

7

Conclusion

7.1 Summary

Folk psychology is, at its core, goal-reason psychology: an understanding of people in terms of the goals they adopt and the reasons they have for adopting these goals and for performing goal-directed actions. It is the consensus view that commonsense goal-reason psychology is belief-desire psychology. According to this BD-Model of folk psychology, our understanding of one another as adopting goals in the light of reasons hinges on the concepts of belief and desire. Essential for proper mastery of the concepts of belief and desire, however, is the capacity to distinguish between the way the world appears to the believer or desirer on the one hand, and the way the world is on the other. Ascription of belief and desire as such must go accompanied by an acknowledgement of the possibility that the ascribed beliefs and desires are or turn out to be false or otherwise inappropriate. Modeling our folk psychological understanding of each other exclusively on these concepts, the BD-Model in effect gives a thoroughly individualist or subjectivist picture of our commonsense conception of the mind. On this picture, the minds of others are essentially private minds, whose subjective, perhaps peculiar and possibly mistaken perspectives on the world are always marked as such and thus distinguished from

the world as we ourselves believe it to be or to become.

This is a distorted picture folk psychology, so I have argued in this book. In its place, I presented the Relational Model of folk psychology. It takes the 'commonsense' in commonsense psychology quite literally, not only in the sense that we rely on a shared sense of discursive understanding of the world in order to engage with one another, but also in the sense that in doing so, we tend to conceive of each other's minds as essentially intersubjective, public phenomena. On this picture, making sense of others is first and foremost an attempt to relate them to a common world, a world shaped by public norms of reason and proper conduct. In this process, we treat their mental states as public states, states consisting in the relation between an individual person and the common world.

On the Relational Model of folk psychology, our basic understanding of other people as rational agents relies on the capacity to draw relations between them and the things in the common world that constitute their goals and reasons. This is what I termed 'relational mindreading' in chapter 2. The kind of social understanding that results from relational mindreading has generally been neglected in the debate on social cognition. The conception of an agent as being genuinely intentionally directed at the world in propositionally articulated ways, without, however, subjectively representing the world in those ways, appears to have been overlooked entirely in the philosophical treatment of folk psychology. With this in mind, the first thing to do was to provide a conceptually coherent story of relational propositional attitudes and their attribution. This was the first challenge I set for the Relational Model in chapter 1 and it was addressed in chapters 3 and 4.

Chapter 3 lead us through Sellars's ingeniously designed Myth of Jones. Jones bootstrapped the Ryleans into a genuine functionalist understanding of mind, of each other's mental states about goals and reasons in particular. But that didn't suffice to introduce them to the idea that people represent their goals and reasons to themselves in ways that may defy public evaluation. Jones taught the Ryleans how to conceive of each other's mental states as FR-representations of the things they intended to achieve and the things that make these ends and their means worth accomplishing. But he didn't show them how to treat such states as *subjective* representations of their goals and reasons, i.e. as full-blown desires and beliefs, respectively. Rylean interpretation was still disjunctively split: An agent acting on beliefs and desires incompatible with public evaluation could not be regarded as behaving in a world-directed way, he either failed in doing so or merely pretended to act in such manner. Exploiting Sellars's verbal behaviorist strategy for our own

expository ends, we thus discovered that the functionalist conception of mind that Sellars introduced in 1956 and that served as the basis for the philosophical treatment of folk psychology in subsequent years, was in fact a relational conception.

Chapter 4 showed that the basic idea of relational mindreading does not depend on any particular conception of the propositional attitudes attributed. The distinction between first-order and second-order mental states enabled us to conceive of relational mindreading outside of the functionalist framework of Sellars's Myth. Relational states may be construed as second-order FR-representational states, but they can also be conceived as first-order states that lack specification in terms of functional or causal role. The differentiation into an intersubjective/public and a subjective/private treatment of mental representation furthermore revealed that no concept of mental representation, functionalist or otherwise, by itself entails the subjective element of our mature understanding of belief and desire. Taken together, these considerations sufficed to demonstrate the conceptual coherence of the notion of relational propositional attitudes and relational mindreading, and thus to establish the conceptual validity of the distinction between (the attribution of) relational and S-representational propositional attitudes.

But is this distinction also empirically robust, does it actually structure our day-to-day discursive engagements with one another? This was the second challenge for the Relational Model and it set the agenda for chapters 5 and 6. Chapter 5 first revealed that the psychological plausibility of relational mindreading is not threatened by dominant explanatory theories found in the current literature, accounts explicitly targeting the cognitive processes of mindreading at the subpersonal level of description. Although clearly inspired by the BD-Model, all these explanatory theories of mindreading turned out to be compatible with a relational interpretation of the explanandum. Two powerful considerations were then presented why such interpretation is actually to be preferred. The BD-Model faces significant problems in accounting for the holistic nature of propositional attitude ascription and the implicit attributions of knowledge that pervade our folk psychological explanations. By taking a different perspective on the explanandum of human discursive understanding, the Relational Model effectively dissolved both problems. These considerations served as input for a discussion of the ontogeny of mindreading. Acquiring a proper understanding of the material (im)proprieties of thought and action in childhood requires the presence of deontic scaffoldings, in the form of caregivers who point out what counts as a proper reason for adopting which goals and performing what kind of actions under which circumstances.

Again, we saw that these considerations were in line with the Relational Model but failed to make proper sense on the BD-Model.

The final hurdle was to account for the complementary roles of belief-desire psychology in discursive practice. This was the focus of chapter 6. Competent use of belief-desire psychology increases our capacity to make sense of seemingly irrational, counter-normative behavior. This facilitates the management of our discursive engagements with one another in a variety of ways. It gives us more interpretative options when seeking to maintain or rehabilitate people's normative status or determining proper sanctions for norm violation. It plays a crucial role in the efficient and effective correction of mistakes and the regulation of future behavior in educative contexts. It moreover enhances our capacity to find compromises in situations of dispute, to protect our own image or to destroy that of others. Finally, S-representational mindreading allows us to develop new standards of objectivity and to critically evaluate, challenge or change the norms that shape common practice.

Together, chapter 5 and 6 completed the second challenge laid out for the Relational Model and established the practical importance of the distinction between relational and S-representational mindreading. Relational mindreading is the central engine of our discursive engagements with one another. It forms the psychological basis on which many of our social practices are built. S-representational mindreading has an essentially complementary function in human interaction. It finds its proper application in relatively complex social situations, situations in which relational mindreading could only lead to misunderstanding.

7.2 The Fallacy Revealed

Why has a relational conception of the propositional attitudes received so little attention in the debate on folk psychology? One of problems I encountered in thinking about folk psychology is that we do not have a distinctive terminology for relational propositional attitudes. In folk psychological practice, factive explanations often suffice to specify the relational attitudes ascribed. John goes to the supermarket in order to buy some milk because he's run out of milk. His relational attitude of desire is conveyed by the description of a future state of affairs *as* his goal, i.e. to buy some milk at the supermarket; his relational attitude of knowledge is presupposed by the description of a fact *as* his reason, i.e. that he has run out of milk. We tend to switch to explicit propositional attitude ascription only when factive explanations no longer suffice,

when there is doubt about the agent's epistemic or motivational state or when generating such explanations would come into conflict with our own commitments as to what counts as an appropriate goal or a reason-constituting fact given the circumstances. Our explicit propositional attitude terminology has a strong S-representational connotation. Using it from a theoretical point of view in the debate on folk psychology can easily make one blind toward the possibility that there might actually be a relational understanding of propositional attitudes underlying our more sophisticated, S-representational concepts.

In chapter 1 I suggested that the BD-Model is the expression of a reflective fallacy: the fallacy of projecting certain philosophical analyses of intentional action in terms of beliefs and desires onto our spontaneous, commonsense understanding of each other in terms of goals and reasons. I also suggested (see chapter 2) that the incentive for the BD-Model comes from a certain understanding of the truth of the following conditional:

> C: Whenever an action is explained in terms of a goal at which it is directed and a reason for which it is performed, there is an explanation of that action in terms of, *inter alia*, a Humean belief-desire pair.

Proponents of the BD-Model want to explain the truth of C either in terms of *conceptual entailment* or in terms of *presupposition* of the consequent by the antecedent. According to the first option, our folk conception of intentional action is such that having a goal and having a reason entails having a desire about that goal and having a belief about that reason. According to the second option, attribution of goals and reasons presupposes not only the truth, existence or expected occurrence of the facts, states of affairs or events that constitute those goals and reasons, but also the presence in the agent of beliefs and desires representing those goals and reasons. In the debate on folk psychology, this second interpretation of conditional C seeps through in the idea that attribution of goals and reasons psychologically requires the ascriptions of goal-representing desires and reason-representing beliefs.

Having met the first challenge outlined above, we can now see that the first option is not available: chapter 3 and 4 showed that there is a conceptual distinction to be drawn between being directed at/representing goals and reasons *simpliciter* on the one hand and *subjectively* representing goals and reasons by instantiating desires and beliefs on the other. Thus, having goals and reasons does *not* entail having subjective desires and beliefs about these

goals and reasons. Having met the second challenge, it also becomes clear why the second option is no longer attractive. Chapter 5 made a strong case for the psychological role of relational mindreading in our everyday discursive understanding of one another. Accordingly, the attribution of goals and reasons in folk psychological practice often need *not* require the ascription of desires and beliefs subjectively representing those goals and reasons; relational mindreading suffices.

In chapter 2.5 I proposed to explain the truth of conditional C not conceptually or psychologically, but *socially* in terms of the rationale of the subjective shift to belief-desire psychology in human discursive practice. The idea was that we can explain the truth of C with reference to a rule in the game of giving and asking for reasons that entitles participants to make the shift towards S-representational mindreading at any stage in the game. Conditional C would then simply be a restatement of that rule: whenever an action can be explained in terms of goals and reasons, *one is entitled* to explain that action in terms of subjective desires and beliefs about those goals and reasons.

After having explored the social functions of belief-desire psychology in chapter 6, we can now explain why there is such a rule. I do not think it is a rule that applies in ordinary folk psychological practice. Explicit reference to subjective beliefs and desires often gives rise to unwanted and inappropriate implicatures. In *philosophical* practice, however, the rule does seem to apply. In chapter 6.4 I argued that belief-desire psychology enables participants in discursive practice to adopt a reflective, critical stance towards common practice. Well, philosophers are most critical folk. They have made it their specialty to adopt a critical attitude toward the relation between mind and world. The truth of C can be explained by pointing out that *within the discursive practice of philosophy*, one is always entitled to make an S-representational shift towards the phenomena under investigation and to study the mind-world relation from a relatively detached, objective point of view.

Conditional C gives expression to a rule in a philosophical language game, a game concerned with providing conceptual clarity to our commonsense understanding of mind and action. But the clarity one aspires should not be measured by the rules that regulate the philosophical practice one is engaged in. The reflective fallacy lies in an attempt to treat a rule of a philosophical game as a logical principle that defines our commonsense concepts, or as a psychological principle that guides our practice of wielding them in quotidian social contexts. It is fed by the intuition that philosophical explanations in terms of beliefs and desires are in some sense *more fundamental* than explanations merely in terms of goals and reasons. But they are not more fundamen-

tal. The philosophical study of mind and action is conceptually and psychologically *more sophisticated* than our mundane ways of thinking about such things, in the sense that it requires more practice and training and might give us better insight. But for all that, it is conceptually and psychologically less fundamental. It only appears to be more fundamental if one mistakes doing philosophy for the real thing.

Bibliography

Altham, J.E.J. 1986. The Legacy of Emotivism. In *Facts, Science and Morality: Essays on A.J. Ayer's Language, Truth and Logic,* ed. G. MacDonald and C. Wright. Oxford: Basil Blackwell.

Alvarez, M. 2010. *Kinds of Reasons.* Oxford: Oxford University Press.

Andrews, K. 2007. Critter Psychology: On the Possibility of Nonhuman Animal Folk Psychology. In *Folk Psychology Re-assessed,* ed. D.D. Hutto and M. Ratcliffe. Dordrecht: Springer.

Andrews, K. 2009. Understanding Norms Without a Theory of Mind. *Inquiry* 52(5): 433-448.

Anscombe, G.E.M. 1957/2000. *Intention.* Cambridge MA: Harvard University Press.

Apperly, I.A. 2011. Mindreaders: *The Cognitive Basis of 'Theory of Mind'.* New York: Psychology Press.

Apperly, I.A. and Butterfill, S.A. 2009. Do humans have two systems to track beliefs and belief-like states? *Psychological Review* 116: 953-70.

Armstrong, D. 1980. *The Nature of Mind and Other Essays.* Brisbane: University of Queensland Press.

Baillargeon, R., Scott, R.M and Zijing, H. 2010. False-belief understanding

in infants. *Trends in Cognitive Science* 14(3): 110-18.

Baron-Cohen, S. 1995. *Mindblindness: An Essay on Autism and Theory of Mind.* Cambridge, MA: MIT Press/Bradford Books.

Baron-Cohen, S., Leslie, A. M. and Frith, U. 1985. Does the autistic child have a "theory of mind"? *Cognition* 21: 37-46.

Baker, L. 1995. *Explaining Attitudes: A Practical Approach to the Mind.* Cambridge: Cambridge University Press.

Baker, L. 1999. What is this Thing Called 'Commonsense Psychology'? *Philosophical Explorations* 1: 3-19.

Bechtel, B. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bermúdez, J.L. 2003. The Domain of Folk Psychology. In *Minds and Persons,* ed. A. O'Hear. Cambridge: Cambridge University Press.

Bermúdez, J.L. 2009. *Philosophy of Psychology: A Contemporary Introduction.* New York: Routledge.

Bittner, R. 2001. *Doing Things for Reasons*. Oxford: Oxford University Press.

Birch, S. and Bloom, P. 2003. Children are Cursed: An Asymmetric Bias in Mental State Attribution. *Psychological Science* 14: 283-86.

Birch, S. and Bloom, P. 2004. Understanding Children's and Adult's Limitations in Mental State Reasoning. *Trends in Cognitive Science* 8: 255-60.

Birch, S. and Bloom, P. 2007. The Curse of Knowledge in Reasoning about False Belief. *Psychological science* 18(5): 382-86.

Borg, E. 2007. If Mirror Neurons are the Answer, What Was the Question? *Journal of Consciousness Studies* 14: 5-19.

Botterril, G. 1996. Folk Psychology and Theoretical Status. In *Theories of Theories of Mind,* ed. P. Carruthers and P. Smith. Cambridge: Cambridge University Press.

Braddon-Mitchell, D. and Jackson, F. 2007. *Philosophy of Mind and Cognition.* Oxford: Blackwell.

Brandom, R.B. 1994. *Making it explicit.* Cambridge, MA: Harvard University Press.

Brandom, R.B. 1997. Study Guide. In *Empiricism and the Philosophy of Mind,* W. Sellars. Cambridge (MA): Harvard University Press.

Brandom, R.B. 2000. *Articulating Reasons: An Introduction to Inferentialism.* Cambridge, MA: Harvard University Press.

Brandom, R.B. 2002. Explanatory vs. Expressive Deflationism about Truth. *In What is Truth?,* ed. R. Schantz. Berlin, New York: Walter de Gruyter.

Bruner, J. 1990. *Acts of Meaning*. Cambridge: Harvard University Press.

Byrne, A. and Logue, H. 2009. *Disjunctivism: Contemporary Readings*. Cambridge, MA: MIT Press.

Carruthers, P. and Smith, P. *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Carruthers, P. 2009. How We Know our own Minds: The Relationship Between Mindreading and Metacognition. *Behavioral and Brain Sciences* 32: 121-82.

Chemero, A. 2010. *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Churchland, P. M. 1981. Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy* 78: 67–90.

Churchland, P. M. 1991. Folk Psychology and the Explanation of Human Behavior. In *The future of Folk Psychology*, ed. J.D. Greenwood. Cambridge: Cambridge University Press.

Csibra, G. and Gergely, G. 2007. Natural Pedagogy. *Trends in Cognitive Science* 13(4): 148-53.

Currie, G. and Sterelny, K. 2000. How to Think About the Modularity of Mindreading. *Philosophical Quarterly* 50: 145-60.

Dancy, J. 2000. *Practical Reality.* Oxford: Oxford University Press.

Dancy, J. 2003. Replies. *Philosophy and Phenomenological Research* 67(2): 468-90.

Davidson, D. 1963. Actions, Reasons, and Causes. *Journal of Philosophy* 60: 685-700. Reprinted in Davidson 2001a.

Davidson, D. 1967. Truth and Meaning. *Synthese* 17: 304-23. Reprinted in Davidson 2001b.

Davidson, D. 1970. Mental Events. In *Experience and Theory,* ed. L. Foster and J.W. Swanson. Amherst: University of Massachusetts Press. Reprinted in Davidson 2001a.

Davidson, D. 1973. Radical Interpretation. *Dialectica* 27: 313-28. Reprinted in Davidson 2001b.

Davidson, D. 1974. Psychology as Philosophy. In *Philosophy of Psychology,* ed. S.C. Brown. The Macmillan Press and Barnes, Noble Inc. Reprinted in Davidson 2001a.

Davidson, D. 1978. Intending. In *Philosophy of History and Action*, ed. Y. Yovel. D. Reidel and The Magnes Press, The Hebrew University. Reprinted in Davidson 2001a.

Davidson, D. 1982. Rational Animals. *Dialectica* 26: 317-27. Reprinted in Davidson 2001c.

Davidson, D. 1983. A Coherence Theory of Truth and Knowledge. In *Kant oder Hegel?*, ed. D. Henrich. Stuttgart: Klett-Cott. Reprinted in Davidson 2001c.

Davidson, D. 1991. Three Varieties of Knowledge. In *A.J. Ayer Memorial Essays: Royal Institute of Philosophy Supplement*, vol. 30, ed. A. Phillips Griffiths. Cambridge, Cambridge University Press. Re printed in Davidson 2001c.

Davidson, D. 1992. The Second Person. In *Midwest Studies in Philosophy, vol. 17*, ed. P. French, T.E. Uehling and W. Wettsein. Indianapolis: University of Notre Dame Press. Reprinted in Davidson 2001c.

Davidson, D. 1999. The Emergence of Thought. *Erkentniss* 51: 7-17. Reprinted in Davidson 2001c.

Davidson, D. 2001a. *Essays on Actions and Events*. New York: Oxford University Press.

Davidson, D. 2001b. *Inquiries into Truth and Interpretation*. New York: Oxford University Press.

Davidson, D. 2001c. *Subjective, Intersubjective, Objective*. New York: Oxford University Press.

De Bruin, L.C. and Strijbos, D.W. 2010. Folk Psychology Without Principles. An Alternative to the Belief-Desire Model of Action Interpretation. *Philosophical Explorations* 13: 257-74.

De Bruin, L.C., Strijbos, D.W. and Slors, M.V.P. 2011. Early Social Cognition: Alternatives to Implicit Mindreading. *Review of Philosophy and Psychology* 2(3): 499-517.

Davies, M. 1994. The Mental Simulation Debate. In *Objectivity, Simulation and the Unity of Consciousness*, ed. C. Peacock. New York: Oxford University Press.

Davies, M. and Stone, T. 1995a. *Folk Psychology.* Cambridge, MA: Blackwell.

Davies, M. and Stone, T. 1995b. *Mental Simulation*. Cambridge, MA: Blackwell.

Dennett, D.C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D.C. 1991. Two Contrasts: Folk-Craft versus Folk-Science, Belief Versus Opinion. In *The Future of Folk Psychology: Intentionality and Cognitive Science*, ed. J. Greenwood. Cambridge: Cambridge University Press.

DeVries, W. 2005. *Wilfrid Sellars*. Montreal: McGill-Queens University Press .

Doherty, M.J. 2009. *Theory of Mind: How Children Understand Others' Thoughts and Feelings*. Hove, UK and New York: Psychology Press.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Flavell, J.H. 1974. *The Development of Inferences about Others. In Understanding Other Persons*, ed. T. Mischel. Oxford: Basil Blackwell.

Flavell, J.H. 1988. The Development of Children's Knowledge About the Mind. From Cognitive Connections to Mental Representations. In *Developing Theories of Mind*, ed. J.W. Astington, P.L. Harris and D.R. Olsen. New York: Cambridge University Press.

Fodor, J. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, J. 1990. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. 1992. A Theory of the Child's Theory of Mind. *Cognition* 44: 283-96. Reprinted in Davies and Stone 1995b.

Fodor, J. 2008. LOT 2: *The Language of Thought Revisited*. Oxford: Oxford University Press.

Frith, U. and Happé, F. 1999. Theory of Mind and Self Consciousness: What Is It Like to Be Autistic? *Mind and Language 14: 1-22*.

Gallagher, S. 2001. The Practice of Mind: Theory, Simulation, or Interaction? *Journal of Consciousness Studies* 8: 83–107.

Gallagher, S. 2008a. Are Minimal Representations Still Representations? *International Journal of Philosophical Studies* 16(3): 351-69.

Gallagher, S. 2008b. Direct Perception in the Intersubjective Context. *Consciousness and Cognition* 17: 535-43.

Gallagher, S. 2011. In Defense of Phenomenological Approaches to Social Cognition: Interacting With the Critics. *Review of Philosophy and Psychology*, DOI: 10.1007/s13164-011-0080-1.

Gallagher, S. and Hutto, D. 2008. Understanding Others Through Primary Interaction and Narrative Practice. In *The Shared Mind: Perspectives on Intersubjectivity*, ed. J. Zlatev, T.P. Racine, C. Sinha and E. Itkonen, E. Amsterdam: John Benjamins.

Gallagher, S. and Zahavi, D. 2008a. *The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science*. London: Routledge.

Gallese, V. 2007. Before and Below 'Theory of Mind': Embodied Simulation and the Neural Correlates of Social Cognition. *Philosophical Transactions of the Royal Society B-Biological Sciences* 362: 659-69.

Gergely, G. and Csibra, G. 2003. Teleological Reasoning in Infancy: The Naïve Theory of Rational Action. *Trends in Cognitive Science* 7(7): 287-92.

Gettier, E. 1963. Is Justified True Belief Knowledge? *Analysis* 23: 121-23.

Gibson. J.J. 1979. *The Ecological Approach to Visual Perception*. Boston:

Houghton Mifflin.

Goldie, P. 2007. There are Reasons and Reasons. In *Folk Psychology Re-assessed*, ed. D.D. Hutto and M. Ratcliffe. Dordrecht: Springer.

Goldman, A. I. 1989. Interpretation Psychologized. *Mind and Language* 4: 161–85. Reprinted in Davies and Stone 1995a.

Goldman, A. I. 1993. The Psychology of Folk Psychology. *Behavioral and Brain Sciences* 16: 15-28.

Goldman, A. I. 2000. Folk Psychology and Mental Concepts. *Protosociology* 14: 102-14.

Goldman, A. I. 2006. *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading*. New York: Oxford University Press.

Gopnik, A. 1996. The Scientist as Child. *Philosophy of Science* 63: 485-514.

Gopnik, A. and Meltzoff, A.N. 1997. *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Gopnik, A. and Wellman, H. 1992. Why the Child's Theory of Mind Really Is a Theory. *Mind and Language* 7, 145-72. Reprinted in Davies and Stone 1995a.

Gopnik, A. and Wellman, H. 1994. The Theory Theory. In *Mapping the Mind: Domain Specificity in Culture and Cognition*, ed. L. Hirschfield and S. Gelman. New York: Cambridge University Press.

Gordon, R.M. 1986. Folk Psychology as Simulation. *Mind and Language* 1: 158-71. In Folk Psychology, ed. M. Davies and T. Stone. Oxford: Blackwell.

Gordon, R.M. 1987. *The Structure of Emotions: Investigations in Cognitive Philosophy*. Cambridge: Cambridge University Press.

Gordon, R.M. 1992. The Simulation Theory: Objections and Misconceptions. *Mind and Language* 7: 11-34. Reprinted in Davies and Stone 1995a.

Gordon, R.M. 1995. Simulation without Introspection or Inference from Me to You. In *Mental Simulation*, ed. M. Davies and T. Stone. Oxford: Blackwell.

Gordon, R. 1996. 'Radical' Simulationism. In *Theories of Theories of Mind*, ed. P. Carruthers and P. Smith. Cambridge: Cambridge University Press.

Gordon, R.M. 2000a. Simulation and the Explanation of Action. In *Empathy and Agency: The Problem of Understanding in the Social Sciences*, ed. B. Koegler and K. Stueber. Oxford: Westview Press.

Gordon, R.M. 2000b. *Sellars' Ryleans Revisited*. Protosociology, 14: 102-14.

Gordon, R. M. 2001. Simulation and Reason Explanation: The Radical View. *Philosophical Topics*, 29: 175-92.

Gordon, R.M. 2005. Intentional Agents Like Myself. In *Perspectives on Imitation: From Cognitive Neuroscience to Social Science: Mechanisms of Imitation and Imitation in Animals (Vol. 1)*, ed. S. Hurley and N. Chater. Cambridge, MA: MIT Press.

Gordon, R.M. 2007. Ascent Routines for Propositional Attitudes. *Synthese* 159: 151-65.

Gordon, R.M. 2008. Beyond Mindreading. *Philosophical Explorations* 11(3): 219-22.

Harris, P., Lillard, A. and Perner J. 1994. Commentary: Triangulating Pretence and Belief. In *Children's Early Understanding of Mind: Origins and Development*, ed. C. Lewis and P. Mitchell. Hove, UK: Lawrence Erlbaum Associates Ltd.

Hamilton, A. 2008. Intention and the Authority of Avowals. *Philosophical Explorations* 11(1): 23-37.

Heal, J. 1986. Replication and Functionalism. In *Language, Mind, and Logic*, ed. J. Butterfield. Cambridge, Cambridge University Press. Reprinted Davies and Stone 1995a.

Heal, J. 1995. How to Think about Thinking. In *Mental Simulation*, ed. M. Davies and T. Stone. Cambridge: Blackwell.

Heal, J. 1996. Simulation, theory and content. In *Theories of Theories of Mind*, ed. P. Carruthers and P. Smith. Cambridge: Cambridge University Press.

Heal. J. 1998: Co-Cognition and Off-line Simulation: Two Ways of Understanding the Simulation Approach. *Mind and Language* 14: 477-98.

Heal, J. 2000. Understanding Other Minds from the Inside. *Protosociology* 14: 102-114.

Heal, J. 2005. Joint Attention and Understanding the Mind. In *Joint Attention: Communication and Other Minds*, ed. N Eilan, C. Horel, T. McCormack and J. Roessler. Oxford: Oxford University Press.

Heil, J. and Mele, A.R. 1993. *Mental Causation*. Oxford: Oxford University Press.

Herschbach, M. 2008a. False-Belief Understanding and the Phenomenological Critics of Folk Psychology. *Journal of Consciousness Studies* 15(12): 33-56.

Herschbach, M. 2008b. Folk Psychological and Phenomenological Accounts of Social Perception. *Philosophical Explorations* 11(3): 223-235.

Hornsby, J. 2008. A Disjunctive Conception of Acting for Reasons. In *Disjunctivism: Perception, Action and Knowledge*, ed. A. Haddock and F. MacPherson. Oxford: Oxford University Press.

Hutto, D.D. 2004. The Limits of Spectatorial Folk Psychology. *Mind and Language* 19: 548–73.

Hutto D.D. 2008a. *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press.

Hutto, D.D. 2008b. The Narrative Practice Hypothesis: Clarifications and Implications. *Philosophical Explorations* 11: 175–92.

Hutto, D.D. 2009. Folk Psychology as Narrative Practice. *Journal of Consciousness Studies*. 16: 9–39.

Hutto, D.D. 2011a. Presumptuous Naturalism: A Cautionary Tale. *American Philosophical Quaterly* 48(2): 129–45.

Hutto, D.D. 2011b. Philosophy of Mind's New Lease on Life: Autopoietic Enactivism meets Teleosemiotics. *Journal of Consciousness Studies* 18(5-6): 44-64.

Hutto, D.D. 2011c. Elementary Mind-Minding, Enactivist-style. In *Joint Attention: New Developments in Psychology, Philosophy of Mind and Social Neuroscience*, ed. A. Seemann. Cambridge, MA: MIT Press.

Hutto, D.D. and Ratcliffe, M. 2007. *Folk Psychology Re-assessed*. Dordrecht: Springer, 25-40.

Jacob, P. 2011. The Direct Perception Model of Empathy: A Critique. *Review of Philosophy and Psychology* 2(3): 515-540.

Jackson F, Pettit P. 1988. Functionalism and Broad Content. *Mind* 97: 381-400.

Jackson, F., Pettit P. 1990. Program Explanation: A General Perspective. *Analysis* 50: 107-17.

Jackson, F., Mason, K. and Stich, S. 2009. Folk Psychology and Tacit Theories: A Correspondence between Frank Jackson and Steve Stich and Kelby Mason. In *Conceptual Analysis and Philosophical Naturalism*, ed. D. Braddon-Mitchell and R. Nola. Cambridge, MA: MIT Press.

Johnson, S. C., Ok, S. and Luo, Y. 2007. The attribution of attention: 9-month-olds' interpretation of gaze as goal-directed action. *Developmental Science* 10(5): 530–37.

Kawada, C.L.K., Oettingen, G., Gollwitzer, P.M. and Bargh, J.A. The Projection of Implicit and Explicit Goals. *Journal of Personality and Social Psychology* 86(4): 445-59.

Keysar, K. and Bly, B. 1995. Intuitions of the Transparency of Idioms: Can One Keep a Secret by Spilling the Beans? *Journal of Memory and Language* 34: 89-109.

Keysar, B., Lin, S. and Barr, D. J. 2003. Limits on Theory of Mind Use in Adults. *Cognition* 89: 25-41.

Kim, J. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Knobe, J.M. 2003. Intentional Action in Folk Psychology: An Experimental Investigation. *Philosophical Psychology* 16(2): 309-24.

Knobe, J.M. 2006. The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology. *Philosophical Studies* 130: 203-31.

Knobe, J.M. and Pettit, P. 2009. The Pervasive Impact on Moral Judgment. *Mind and Language* 24(5): 586-604.

Leslie, A.M. 1987. Pretense and Representation: The Origins of "Theory of Mind". *Psychological Review* 94: 412–426.

Leslie, A. 1994. ToMM, ToBy, and Agency: Core Architecture and Domain Aspecificity. In *Mapping the Mind: Domain Specificity in Cognition and Culture*, ed. L. Hirschfeld and S. Gelman. Cambridge: Cambridge University Press.

Leslie, A.M. 2000. 'Theory of mind' as a Mechanism of Selective Attention. In *The New Cognitive Neurosciences*, ed. M. Gazzaniga. Cambridge, MA: MIT Press.

Leslie, A. 2005. Developmental Parallels in Understanding Minds and Bodies. *Trends in Cognitive Science* 9: 459–62.

Leslie, A.M. and Polizzi, P. 1998. Inhibitory Processing in the False Belief Task: Two Conjectures. *Developmental Science* 1: 247–53.

Leslie, A.M., German, T. and Polizzi, P. 2005. Belief-Desire Reasoning as a Process of Selection. *Cognitive Psychology* 50: 45-85.

Lewis, D. 1970. How to Define Theoretical Terms. *Journal of Philosophy* 67: 427-46.

Lewis, D. 1972. Psychophysical and Theoretical Identifications. *Australian Journal of Philosophy* 50:249-58.

Lewis, D. 1979. Attitudes De Dicto and De Se. *The Philosophical Review* 88(4): 513-43.

Luo, Y. and Baillargeon, R. 2007. Do 12.5-Month-Old Infants Consider What Objects Others Can See when Interpreting Their Actions? *Cognition* 105: 489-512.

Luo, Y. and Beck. W. 2010. Do You See What I See? Infants' Reasoning About Others' Incomplete Perceptions. *Developmental Science*

13: 134-42.

Malle, B.F. 2001. Folk Explanations of Intentional Action. In *Intentions and Intentionality: Foundations of Social Cognition*, ed. B.F. Malle, L.J. Moses, and D.A. Baldwin. Cambridge MA: MIT Press.

Malle, B.F. 2004. *How The Mind Explains Behavior: Folk Explanations, Meaning and Social Interaction*. Cambridge, MA: MIT Press.

Malle, B.F., Knobe, J.M. and Nelson, S.E. 2007. Actor-Observer Asymmetries in Explanations of Behavior: New Answers to an Old Question. *Journal of Personality and Social Psychology* 93(4): 491-514.

McDowell, J. 1978. Are Morel Requirements Hypothetical Imperatives? *Proceedings of the Aristotelian Society* 52: 13-29. Reprinted in McDowell 1998.

McDowell, J. 1994. *Mind and World*. Cambridge MA: Harvard University Press

McDowell, J. 1998. *Mind, Value and Reality*. Cambridge, MA: Harvard University Press.

McDowell, J. 1998. Having the World in View: Sellars, Kant and Intentionality. *The Journal of Philosophy* 95(9): 431-491. Reprinted in McDowell 2009.

McDowell, J. 2009. *Having the World in View.* Cambridge MA: Harvard University Press.

McGeer, V. 2001. Psycho-practice, Psycho-theory and the Contrastive Case of Autism. How Practices of Mind Become Second-Nature. *Journal of Consciousness Studies* 8(5–7): 109–32.

McGeer, V. 2007. The Regulative Dimension of Folk Psychology. In *Folk Psychology Re-assessed*, ed. D.D. Hutto and M. Ratcliffe. Dordrecht: Springer.

Menary, R. 2006. *Consciousness and Emotion: A Special Issue on Radical Enactivism*. Philadelphia: John Benjamins.

Millikan, R. G. 1984. *Language, Thought and other Biological Categories*. Cambridge: MIT Press.

Millikan, R.G. 2004. *Varieties of Meaning*. Cambridge (MA): MIT Press.

Millikan, R. G. 2005. The Father, the Son, and the Daughter: Sellars, Brandom, and Millikan. *Pragmatics and cognition* 13(1): 59–72.

Moore, C., Jarrold, C., Russell, J., Lumb, A., Sapp, F. and MacCallum, F. 1995. Conflicting Desires and the Child's Theory of Mind. *Cognitive Development* 49: 467-482.

Morton, A. 1980. *Frames of Mind: Constraints on the Common Sense Conception of the Menta*l. Oxford: Oxford University Press.

Morton, A. 1996. Folk Psychology Is Not a Predictive Device. *Mind* 105(417): 119-37.

Morton, A. 2003. *The Importance of Being Understood: Folk Psychology as Ethics*. London: Routledge.

Nickerson, R.S. 1999: How We Know – and Sometimes Misjudge – What Others Know: Imputing One's Own Knowledge to Others. *Psychological Bulletin* 125: 737–59.

Nickerson, R.S., 2001: The Projective Way of Knowing: A Useful Heuristic That Sometimes Misleads. *Current Directions in Psychological Science* 10: 168-72.

Nichols, S. and Stich, S. 1998. Rethinking Co-cognition: A Reply to Heal. *Mind and Language* 13(4): 499-512.

Nichols, S. and Stich, S. 2003. *Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding of Other minds*. Oxford: Clarendon Press.

O'Brien, G. and Opie, J. 2004. Notes Toward a Structuralist Theory of Mental Representation. In *Representation in Mind: New Approaches to Mental Representation*, ed. H. Clapin, P. Staines and P. Slezak. Oxford: Elsevier Ltd.

O'Brien, G. and Opie, J. 2011. Representation in Analog Computation. In *Knowledge and Representation*, ed. A. Newen, A. Bartels and E. Jung. CSLI Publications.

Onishi, K.H. and Baillargeon, R. 2005. Do 15-Month-Old Infants Understand False Beliefs? *Science* 308: 255-58.

O' Shea, J.R. 2007. *Wilfrid Sellars*. Cambridge: Polity Press.

Parsell, M 2010. Sellars on Thoughts and Beliefs. *Phenomenology and the Cognitive Sciences* 10(2): 261-75.

Perner, J. 1988. Developing Semantics for Theories of Mind. From Propositional Attitudes to Mental Representation. In: *Developing Theories of Mind*, ed. J.W. Astington, P.L. Harris, D.R. Olsen. New York: Cambridge University Press.

Perner, J. 1991. *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Perner, J. 1995. The Many Faces of Belief: Reflections on Fodor's and the Child's Theory of Mind. *Cognition* 57: 241-69.

Perner, J. 2010. Who took the Cog out of Cognitive Science? Mentalism in an Era of Anti-cognitivism. *Cognition and Neuropsychology International Perspectives on Psychological Science (Volume 1)*, ed. P. A. Frensch and R. Schwarzer. New York: Psychology Press.

Perner, J., Baker, S. and Hutton, D. 1994. Prelief: The Conceptual Origins of Belief and Pretence. In *Children's Early Understanding of Mind: Origins and Development,* ed. C. Lewis and P. Mitchell. Hove, UK: Lawrence Erlbaum Associates Ltd.

Perner, J., Stummer, S., Sprung, M. and Doherty, M. 2002. Theory of Mind Finds its Piagetian Perspective: Why Alternative Naming Comes with Understanding Belief. *Cognitive Development* 17: 1451–72.

Perner, J., Brandl, J. and Garnham, A. 2003. What is a Perspective Problem? Developmental Issues in Belief Ascription and Dual Identity. *Facta Philosophica* 5: 355-78.

Perner, J. and Ruffman, T. 2005. Infants' Insight into the Mind: How Deep? *Science* 308: 214–16.

Perner. J, Zauner, P. and Sprung, M. 2005. What Does "That" Have to Do with Point of View? In *Why Language Matters for Theory of Mind,* ed. J.W. Astington and J.A. Baird. Oxford: Oxford University Press.

Perner, J. and Roessler, J. 2010. Teleology and Causal Understanding in Children's Theory of Mind. In *Causing human Actions: New Perspectives on the Causal Theory of Action*, ed. J.H. Aguilar and A.A. Buckareff. Cambridge MA: MIT Press.

Povinelli, D.J. and Vonk, J. 2003. Chimpanzees Minds: Suspiciously Human? *Trends in Cognitive Sciences* 7: 157-60.

Povinelli, D.J. and Vonk, J. 2004. We Don't Need a Microscope to Explore the Chimpanzee's Mind. *Mind and Language* 19(1): 1-28.

Premack, D. and Woodruff, G. 1978. Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences* 1: 515-26.

Pylyshyn, Z.W. 1978. When is the Attribution of Beliefs Justified? *The Behavioral and Brain Sciences* 1: 592-93.

Quine, W.V.O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Quine, W.V.O. 1969. Propositional Objects. In *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Rakoczy, H. 2010. Executive Function and the Development of Belief–Desire Psychology. *Developmental Science* 13:4: 648–61.

Rakoczy, H., Warneken, F. and Tomasello, M. 2007. "This way!", "No! That way!"—3-Year Olds Know that Two People Can Have Mutually Incompatible Desires. *Cognitive Development* 22: 47–68.

Rakoczy, H., Warneken, F. and Tomasello, M. 2008. The Sources of Normativity: Young Children's Awareness of the Normative Structure of Games. *Developmental Psychology* 44(3): 875–81.

Rakoczy, H., Warneken, F. and Tomasello, M. 2009. Young Children's

Selective Learning of Rule Games From Reliable and Unreliable Models. *Cognitive Development* 24(1): 61-69.

Ramsey, W. M. 2007. *Representation Reconsidered.* New York: Cambridge University Press.

Ramsey, W., S. Stich and J. Garon. 1991. Connectionism, Eliminativism and the Future of Folk Psychology. In *The Future of Folk Psychology,* ed. J. Greenwood. Cambridge: Cambridge University Press.

Ratcliffe, M. 2006. 'Folk Psychology' is not Folk Psychology. *Phenomenology and the Cognitive Sciences* 5:31-52.

Ratcliffe, M. 2007. *Rethinking Commonsense Psychology: A Critique of Folk Psychology, Theory of Mind and Simulation*. Basingstoke: Palgrave Macmillan.

Ratcliffe, M. 2009. There are no Folk Psychological Narratives. *Journal of Consciousness Studies* 16(6–8): 379–406.

Ravenscroft, I. 2003. Simulation, Collapse and Humean Motivation. *Mind and Language* 18(2): 162-74.

Repacholi, B. M. and Gopnik, A. 1997. Early Reasoning about Desires: Evidence from 14- and 18-Month Olds. *Developmental Psychology* 33: 12-21.

Rosenberg, J. F. 2004. Ryleans and outlookers: Wilfrid Sellars on "Mental States". *Midwest Studies in Philosophy* 28: 239–65. Reprinted in Rosenberg 2007.

Rosenberg, J.F. 2007. *Wilfrid Sellars: Fusing the Images*. Oxford: Oxford University Press.

Ruffman, T. and Perner, J. 2005. Do Infants Really Understand False Belief? Response to Leslie. *Trends in Cognitive Science* 9: 462–63.
Ryle, G. 1949. *The concept of mind*. London: Hutchison.

Scharp, K. and Brandom R.B. 2007. *In the Space of Reasons: Selected Essays of Wilfrid Sellars*. Cambridge, MA: Harvard University Press.

Schueler, G.F. 2003. *Reasons and Purposes. Human Rationality and the Teleological Explanation of Action.* Oxford: Oxford University Press.

Schwitzgebel, E. 2002. A Phenomenal, Dispositional Account of Belief. *Noûs* 36(2): 249-75.

Scott, R.M., and Baillargeon, R. 2009. Which Penguin Is This? Attributing False Beliefs about Object Identity at 18 Months. *Child Development* 80: 1172–96.

Searle, J. 1983. *Intentionality, an Essay in the Philosophy of Mind*. New

York: Cambridge University Press.
Sellars, W. 1953. Inference and Meaning. *Mind* 62: 313-38. Reprinted in Scharp and Brandom 2007.
Sellars, W. 1954. Some Reflections on Language Games. *Philosophy of Science* 21: 204-28. Reprinted in Scharp and Brandom 2007.
Sellars, W. 1956/1997. *Empiricism and the Philosophy of Mind*. Cambridge (MA): Harvard University Press.
Sellars, W. 1963/1991. *Science, Perception and Reality*. Atascedero, CA: Ridgeview Publishing Co.
Sellars, W. 1966. Thought and Action. In *Freedom and Determinism*, ed. K. Lehrer. New York: Random House.
Sellars, W. 1967/1992. *Science and Metaphysics: Variations on Kantian Themes*. Atascedero, CA: Ridgeview Publishing Co.
Sellars, W. 1969. Language as Thought and as Communication. *Philosophy and Phenomenological Research* 29: 506-27. Reprinted in Scharp and Brandom 2007.
Sellars, W. 1973. Actions and Events. *Noûs* 7: 179–202.
Sellars, W. 1974. Meaning as functional classification. *Synthese* 27: 417-70. Reprinted in Scharp and Brandom 2007.
Sellars, W. 1980. *Naturalism and Ontology*. Atascedero, CA: Ridgeview Publishing Co.
Surian, L., Caldi, S. and Sperber, D. 2007. Attribution of Beliefs to 13-Month-Old Infants. *Psychological Science* 18: 580-86.
Scholl, B. J. and Leslie, A. M. 1999. Modularity, Development and 'Theory of Mind'. *Mind and Language* 14: 131-53.
Smith, M. 1987. The Humean Theory of Motivation. *Mind* 96: 36-61.
Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell Publishing.
Smith, M. 1998. The Possibility of Philosophy of Action. In *Human Action, Deliberation and Causation*, ed. J. Bransen en S. Cuypers. Dordrecht: Kluwer Academic Publishers. Reprinted in Smith 2004.
Smith, M. 2003. Humeanism, Psychologism and the Normative Story. *Philosophy and Phenomenological Research* 67(2): 460-67. Reprinted in Smith 2004.
Smith, M. 2004. *Ethics and the A Priori*. Cambridge: Cambridge University Press.
Southgate, V., Senju, A. and Csibra, G. 2007. Action Anticipation through Attribution of False Belief by Two-Year-Olds. *Psychological Science* 18: 587-92.
Spaulding, S. 2010. Embodied Cognition and Mindreading. *Mind and*

Language 25(1): 119-40.

Stich, S. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: MIT Press.

Stich, S. and Nichols, S. 1992. Folk Psychology: Simulation or Tacit Theory? *Mind and Language* 7: 35-71. Reprinted in Davies and Stone 1995a.

Stich, S. and Nichols, S. 1995. Second Thoughts on Simulation. In *Mental Simulation*, ed. M Davies and T Stone. Oxford: Blackwell.

Stich, S. and Nichols, S. 1997. Cognitive Penetrability, Rationality and Restricted Simulation. *Mind and Language* 12(3-4): 297-326.

Stich, S. and Ravenscroft, I. 1994. What is Folk Psychology? *Cognition* 50: 447-68.

Stone, T. and Davies, M. 2001. Mental Simulation, Tacit Theory, and the Threat of Collapse. *Philosophical Topics* 29(1-2): 127-73.

Stoutland, F. 2007. Reasons for Action and Psychological States. In *Action in Context*, ed. A. Leist. Berlin: Walter de Gruyter GmbH & Co.

Strawson, P.F. 1962. Freedom and Resentment. *Proceedings of the British Academy* 48: 1-25.

Strijbos, D.W. and De Bruin, L.C. 2012. Making Folk Psychology Explicit: The Relevance of Robert Brandom's Philosophy for the Debate on Social Cognition. *Philosophia* 40(1): 139-63.

Surian, L., Caldi S, and Sperber, D. 2007. Attribution of Beliefs to 13-Month-Old Infants. *Psychological Science* 18: 580-86.

Thompson, E. 2007. *Mind in Life. Biology, Phenomenology and the Science of Mind*. Cambridge, MA: Harvard University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T. and Moll, H. 2005. Understanding and Sharing Intentions: The Origins of Cultural Cognition. *Behavioral and Brain Sciences* 28: 675-91.

Van Boven, L., Dunning, D. and Loewenstein, G. 2000. Egocentric Empathy Gaps Between Owners and Buyers. *Journal of Personality and Social Psychology* 79: 66-76.

Van Boven, L. and Loewenstein, G. 2003. Social Projection of Transient Drive States. *Personality and Social Psychology Bulletin* 29: 1159-68.

Weiskopf, D. 2005. Mental Mirroring as the Origin of Attributions. *Mind and Language* 20: 495-520.

Wellman, H.M., Cross, D and Watson, J. 2001. Meta-analysis of Theory of Mind Development: The Truth about False Belief. *Child Development* 72: 655-84.

Wilkerson, W.S. 2001. Simulation, Theory, and the Frame Problem. The Interpretative Moment. *Philosophical Psychology* 14(2): 141-53.

Wilkes, K. 1991. The Relationship between Scientific Psychology and Common-Sense Psychology. *Synthese* 89: 15-39.

Williams, B.A.O. 1981. *Moral luck*. Cambridge: Cambridge University Press.

Williamson, T. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

Wimmer, H. and Perner, J. 1983. Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition* 13, 103-28.

Wittgenstein, L. 1953/2001. *Philosophical Investigations*. Oxford: Blackwell Publishing.

Woodward, A. L. 2003. Infants' developing understanding of the link between looker and object. *Developmental Science* 6(3): 297–311.

Woodward, A.L. 2005. Infants' Understanding of The Actions Involved in Joint Attention. In *Joint Attention: Communication and Other Minds*, ed. N. Eilan, C. Horel, T. McCormack and J. Roessler. Oxford: Oxford University Press.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Yuill, N., Perner, J., Peerbhoy, D and Van den Ende, J. 1996. Children's Changing Understanding of Wicked Desires: From Objective to Subjective Moral. *British Journal of Developmental Psychology* 14: 457-475.

Zahavi, D. 2007. Expression and Empathy. In *Folk Psychology Re-assessed*, ed. D.D. Hutto, M. Ratcliffe. Dordrecht: Springer.

Zahavi, D. 2011. Empathy and Direct Social Perception: A Phenomenological Proposal. *Review of Philosophy and Psychology* 2(3): 541-558.

Zahavi, D. and Gallagher, S. 2008. The (In)visibility of Others: A Reply to Herschbach. *Philosophical Explorations* 11(3): 237-44.

Zawidzki, T. W. 2008. The Function of Folk-Psychology: Mind-Reading or Mind-Shaping? *Philosophical Explorations* 11(3): 193-17.

Zawidzki, T. W. 2011. How to Interpret Infant Socio-Cognitive Competence. *Review of Philosophy and Psychology* 2(3): 483-97.

Zawidzki, T. W. Forthcoming. *Mindshaping: Linchpin of the Human Socio-Cognitive Syndrome*. MIT-Press.

Samenvatting (Summary in Dutch)

'Folk psychology' is de psychologie die we in de alledaagse sociale praktijk gebruiken om elkaar en onszelf te begrijpen. Het omvat een rijke verzameling van begrippen die allemaal te maken hebben met onze psyche, geest of 'mind', begrippen die we nodig hebben om te begrijpen waarom iemand iets denkt, voelt, zegt, doet of juist nalaat. Dit proefschrift spitst zich toe op de psychologie die we doorgaans gebruiken om elkaars gedrag te begrijpen als *intentionele handelingen*, d.w.z. als gedrag met *doelen* en *redenen*.

Stel je ziet je partner 's ochtends na het ontbijt op een doordeweekse dag haastig zijn jaszakken doorzoeken. Je interpreteert dit gedrag als een handeling die hij uitvoert met het *doel* zijn autosleutels te vinden. Je partner doet dit natuurlijk niet zomaar. Hij zoekt zijn autosleutels *omdat* het nu echt tijd is om naar zijn werk te gaan (zijn *reden* om naar zijn autosleutels te zoeken). Het interpreteren van elkaars gedrag in termen van doelen en redenen voor handelen is iets wat we voortdurend doen, meestal zonder erbij na te denken. Het is een sociaal-cognitief vermogen dat vaak zo vanzelfsprekend is, dat je bijna zou vergeten hoe belangrijk het is voor het dagelijkse intermenselijke verkeer. Zonder folk psychology zou je niet in staat zijn het simpele scenario hierboven te begrijpen. Je zou geen idee hebben in welke context je de graaiende handbewegingen van je partner moest plaatsen. Folk psychology ligt ten grondslag aan alle complexe sociale relaties en omgangsvormen die onze maatschappij

typeren. Het is een begrippenkader dat elk zich normaal ontwikkeld kind zich toe-eigent en dat het nodig heeft om volwaardig deel te kunnen nemen aan onze samenleving.

Folk psychology, in engere zin, heeft dus betrekking op onze sociale vaardigheid elkaars gedrag te interpreteren in termen van doelen en redenen voor handelen. Kort gezegd: op ons vermogen te snappen wat iemand doet en waarom. In dit proefschrift beschrijf ik dit vermogen als een typisch rationele, *discursieve* vorm van sociale cognitie. Door gebruik te maken van folk psychology plaatsen we gedrag in een redelijke context. Dit stelt ons in staat om met elkaar in gesprek te komen over wat we hebben gedaan of willen gaan doen, om extra uitleg te geven over het hoe en waarom of verdere afwegingen te maken over wat het meest wenselijk of praktisch is, gegeven de omstandigheden. In de filosofische en wetenschappelijke literatuur wordt ook wel gesproken over 'mentaliseren' of 'mindreading' wanneer men het heeft over deze discursieve vorm van sociale cognitie. Het betreft de vaardigheid andermans gedrag te interpreteren in de context van zijn of haar mentale leefwereld, om iemands 'mind' te kunnen 'lezen' in het gedrag dat hij of zij vertoont in een bepaalde situatie.

In de filosofische en wetenschappelijke literatuur over folk psychology houdt men zich met name bezig met de vraag hoe deze vorm van sociale cognitie psychologisch of neurobiologisch verklaard moet worden. Is folk psychology bijvoorbeeld een *theorie* met behulp waarvan we interne causaal effectieve toestanden postuleren achter het gedrag dat we proberen te begrijpen, analoog aan de manier waarop wetenschappers (cognitief psychologen, neurowetenschappers) gedragsfenomenen proberen te verklaren? Of is mentaliseren veeleer een vorm van *simulatie* die ons in staat stelt ons te verplaatsen in het perspectief van de ander om zodoende te ervaren wat hij of zij van plan is en waarom?

De discussie tussen aanhangers van deze twee stromingen (de zogenaamde 'theorie theorie' en 'simulatie theorie') heeft het filosofische en wetenschappelijke debat over folk psychology de afgelopen decennia sterk bepaald. Met een focus op het vinden van een psychologische of neurobiologische *verklaring* van ons vermogen tot mentaliseren is een behoorlijke discussie over een adequate *karakterisering* ervan echter achterwege gebleven. Aanhangers van zowel de theorie theorie als de simulatie theorie veronderstellen veelal een begrip van mentaliseren dat, zo beweer ik in dit proefschrift, ongenuanceerd en onvolledig is en daarom geen goede afspiegeling kan vormen van het fenomeen dat ze proberen te verklaren. Door de exclusieve focus op verklarende theorieën is een belangrijk *explanandum* (wat verklaard moet worden) van dis-

cursieve sociale cognitie over het hoofd gezien. Het doel van dit proefschrift is om deze vorm van doel-redeninterpretatie voor het voetlicht te brengen en te laten zien welke cruciale rol het speelt in onze dagelijkse sociale praktijk.

Een ander voorbeeld. Stel een vriendin belt je op om af te zeggen voor vanavond. Haar moeder is plotseling ernstig ziek geworden en ze zit nu in de auto op weg naar haar moeder om haar bij te staan. De uitleg van je vriendin kan als volgt worden weergegeven.

U1: V is op weg naar haar moeder…

- om haar bij te staan
- omdat zij plotseling ernstig ziek is geworden.

Het eerste gedeelte van de uitleg geeft V's doel weer: ze is op weg naar haar moeder *met als doel* haar bij te staan. Het tweede gedeelte laat zien waarom V dit doel heeft: V wil haar moeder bij staan *met als reden* dat haar moeder plotseling ernstig ziek is geworden. Zo op het eerste gezicht geeft U1 een prima verklaring van V's gedrag. En in principe heb je ook geen extra informatie nodig om te snappen wat ze aan het doen is en waarom. Merk echter op dat U1 geen expliciete mentaliserende terminologie bevat. V verklaart haar gedrag met verwijzing naar haar doel (haar moeder bijstaan) en haar reden daarvoor (haar moeder is plotseling ernstig ziek geworden). Het doel van haar handeling wordt hier verwoord als een toekomstige gebeurtenis (haar moeder bijstaan) de reden als een feit of stand van zaken (haar moeder is plotseling ernstig ziek geworden). V verklaart haar gedrag door te verwijzen naar gebeurtenissen, feiten of standen van zaken in de buitenwereld, niet door expliciet te refereren aan haar eigen mentale toestanden.

Hoe moeten we dit nu begrijpen? Is V's uitleg bij nader inzien dan toch geen vorm van mentaliseren, maar slechts een soort van reductionistische stimulus-responsverklaring van gedrag in termen van 'input' vanuit de buitenwereld? Een dergelijk radicaal behaviorisme is niet erg plausibel. Immers, V's uitleg laat duidelijk zien dat haar gedrag een intentionele handeling is met doelen en redenen. En een intentionele handeling veronderstelt een mind die gedachten heeft, plannen maakt en intenties vormt: mentale activiteit die niet tot gedrag is te reduceren. Wat maakt V's uitleg dan tot een mentaliserende interpretatie? In het debat over folk psychology luidt het standaard antwoord op deze vraag als volgt: je interpreteert V's uitleg in termen van doelen en redenen omdat je haar bewust of onbewust mentale toestanden toeschrijft die haar doelen en redenen *representeren*. Discursieve sociale cognitie, zo luidt de

consensus, vindt haar essentie dus in het toeschrijven van representationele mentale toestanden.

Ik denk dat dit een verkeerde manier is om ons alledaags begrip van uitspraken als U1 te karakteriseren. De stelling van dit proefschrift is dat discursieve sociale cognitie doorgaans niet representationeel is, maar *relationeel*. Dit betekent grofweg dat we de minds van anderen, door hun gedrag te *relateren aan* de buitenwereld, begrijpen *in termen van* de buitenwereld. Maar dit maakt het gedrag van anderen niet 'mindless', zoals in een radicaal-behavioristische verklaring, het maakt de buitenwereld juist 'mindful'.

Volgens het standaard model van folk psychology maken wij in ons alledaags begrip van uitspraken als U1 noodzakelijkerwijs gebruik van de concepten van 'belief' en 'desire' (enigszins krom vertaald als 'overtuiging' of 'geloof' en 'verlangen'). Het zogenaamde 'Belief-Desire Model' (BD-Model) van folk psychology stelt dat U1 per definitie onvolledig is en slechts een verkorte versie is van de *werkelijke* verklaring:

U2: V is op weg naar haar moeder…

- met *het verlangen* haar bij te staan
- omdat *ze gelooft dat* zij plotseling ernstig ziek is geworden.

Het BD-Model zegt hiermee dat het toeschrijven van doelen en redenen voor handelen de begrippen van 'belief' en 'desire' impliciet veronderstelt: verlangens (desires) naar de toegeschreven *doelen* en overtuigingen (beliefs) over de toegeschreven *redenen*. Beliefs en desires worden in de filosofische traditie als prototypes beschouwd van *representationele, subjectieve* mentale toestanden. Ze geven uitdrukking aan de manier waarop degene aan wie ze toegeschreven worden, de wereld representeert, hoe deze persoon de wereld subjectief ervaart. Dit komt het meest duidelijk naar voren wanneer je iemand een *onjuiste* overtuiging toeschrijft: je begrijpt deze persoon dan als iemand die de wereld mis-representeert, iemand die een subjectief oordeel heeft over de wereld dat niet strookt met de werkelijkheid. Iets dergelijks doet zich ook voor wanneer je iemands gedrag interpreteert in termen van een *onrealistisch* of *ongepast* verlangen. De persoon die zich een doel stelt vanuit een onrealistische verlangen is iemand wiens subjectieve verwachting ten aanzien van de toekomst niet reëel is, iemand die de toekomst dus op een bepaalde manier mis-representeert. En een ongepast verlangen representeert een doel dat iemand niet zou moeten hebben in de specifieke situatie waarin hij of zij verkeert: het is een mis-rep-

resentatie van de praktische normen die gelden in de betreffende situatie.[106]

Een goed verstaan van de begrippen 'belief' en 'desire' vereist dus dat we een onderscheid maken tussen de wereld zoals hij is (zoals hij ons toeschijnt) en de wijze waarop degene aan wie beliefs en desires wordt toegeschreven de wereld subjectief representeert. Volgens het BD-Model ontleent elke uitleg van andermans handelen in termen van gebeurtenissen en feiten of standen van zaken in de buitenwereld, zijn verklarende kracht aan een onderliggende representationele verklaring in termen van beliefs en desires. Het BD-Model schetst een exclusief subjectivistisch beeld van de discursieve mind. Het veronderstelt daarmee een vorm van *mind-wereld dualisme* dat mijns inziens ons spontane begrip van elkaar als gesituationeerde, rationele actoren wezensvreemd is.

Het alternatief dat ik in dit proefschrift hiertegenover plaats, begint bij de simpele observatie dat in de meeste alledaagse situaties een representationeel begrip van elkaars handelen helemaal niet nodig is. In verreweg de meeste gevallen delen we onze overtuigingen en verlangens ten aanzien van redenen en doelen voor handelen in een bepaalde praktische situatie. Als je met je auto naar werk moet, ga je de autosleutels zoeken. Bijvoorbeeld in de zakken van je jas. En wanneer je hoort dat je moeder plotseling ernstig ziek is geworden, ga je naar haar toe om haar bij te staan. Natuurlijk geldt dit niet voor iedereen in elke situatie, maar dit zijn de uitzonderingen die de regel bevestigen.

Over het algemeen handelen we zoals we behoren te handelen, met de doelen en redenen die we behoren te hebben, gegeven de specifieke omstandigheden. En dit is maar goed ook. Mensen die in hun gedrag te veel afwijken van geldende socio-culturele normen, worden onvoorspelbaar. En met onvoorspelbare mensen kun je niet samenwerken, laat staan samenleven. Folk psychology is bruikbaar omdat en voor zover we handelen overeenkomstig

106 In de ontwikkelingspsychologie wordt het begrip van beliefs en desires bij jonge kinderen getest met behulp van zgn. 'false belief tests' en 'discongruent desire tests'. Bij de klassieke false belief test kijkt een kind naar een scenario met twee poppen, waarna het een voorspelling moet doen van het gedrag van een van de poppen. Om de juiste voorspelling te geven, zo is de redenering, moet het kind een false belief aan de betreffende pop toeschrijven. Het scenario is bijvoorbeeld als volgt. Pop A stopt een voorwerp X in een mand en plaatst vervolgens het deksel weer op de mand, zodat het voorwerp niet meer te zien is. Naast de mand staat een kast met een deur. Na voorwerp X in de mand te hebben gestopt verlaat pop A de ruimte. Vervolgens komt pop B binnen, haalt X uit de mand, plaatst het deksel weer op de mand, stopt X in de kast, en sluit de kastdeur, zodat ook nu het voorwerp niet te zien is. Hierna verlaat pop B de ruimte. Wanneer pop A weer terugkomt, wordt aan het kind de vraag gesteld waar A zal gaan kijken om X te pakken: in de mand of in de kast? Als het kind het goede antwoord geeft (de mand), zou het snappen dat A een subjectieve mis-representatie heeft van voorwerp X als zijnde in de mand, terwijl X in de kast ligt. Bij dergelijke klassieke false belief tests geven kinderen pas vanaf ongeveer 4-jarige leeftijd het goede antwoord. Jongere kinderen wijzen stelselmatig naar de kast wanneer hen de vraag gesteld wordt. De verschillende false belief tests en discongruent desire tests worden besproken in de appendix bij hoofdstuk 5.

de normen die folk psychology veronderstelt. Een goede opvoeding en gedegen onderwijs zorgt er in de meeste gevallen voor dat we ons aan de normen houden en een praktische situatie inschatten zoals we dat behoren te doen, door de juiste doelen te stellen op basis van de juiste redenen, gegeven die situatie. Dit *normatieve* aspect van folk psychology wordt nogal eens over het hoofd gezien. We maken gebruik *van* folk psychology om anderen te kunnen mentaliseren. Maar mentaliseren of mindreading zou praktisch onmogelijk zijn zonder een pedagogisch proces van *mindshaping*, waarin onze mindset gevormd wordt door de gedragsnormen de impliciet gesteld worden *door* folk psychology.

Zolang anderen handelen in overeenstemming met hoe het hoort, is een representationeel begrip van hun mind niet nodig. Als je niet precies begrijpt wat ze aan het doen zijn, volstaat informatie over de gebeurtenissen, feiten of standen van zaken in de buitenwereld die hun doelen en redenen vormen. Het heeft dan geen praktisch toegevoegde waarde om stil te staan, bewust of onbewust, bij hun *subjectieve representatie van* deze gebeurtenissen, feiten of standen van zaken. Een representationeel begrip van rationeel handelen wordt van belang wanneer anderen praktische of epistemische normen lijken te schenden, wanneer er onduidelijkheid is over deze normen of wanneer we het van belang achten ze te evalueren of te veranderen.

Het alternatief dat ik tegenover het BD-Model plaats, noem ik in hoofdstuk 2 het 'Relationele Model' van folk psychology. Het Relationele Model onderscheidt twee vormen van discursieve sociale cognitie: 'relational mindreading' en 'representational mindreading'. Representational mindreading is belief-desire psychologie, zoals beschreven door het BD-Model. Het Relationele Model stelt echter dat representational mindreading een typisch *reflectieve* manier is van interpretatie, een uitzonderlijke vorm die we nodig hebben in problematische sociale situaties, zoals hierboven beschreven: wanneer we normen lijken te schenden, etc.

De cognitieve basis van onze dagelijkse discursieve interactie bestaat echter uit relational mindreading. Relational mindreading is onze *spontane* manier van discursieve sociale cognitie. Het is een echte vorm van mentaliseren, maar zonder de ander te conceptualiseren als iemand die de wereld subjectief representeert. In de act van relational mindreading relateer je de ander aan de doelen en redenen in de buitenwereld die zijn of haar handeling begrijpelijk maken. Dit betekent dat je op dat moment geen rekening houdt met de mogelijkheid dat de doelen en redenen van de ander inadequaat (onhaalbaar, ongepast, niet gebaseerd op feiten, etc.) zijn. Je gaat ervan uit dat de ander handelt zoals hij zou moeten handelen in zijn situatie. De technische term

'relationeel' is gekozen omdat het veronderstelt dat beide zgn. 'relata' (wat aan elkaar gerelateerd wordt) in de wereld aanwezig zijn (waren of zullen zijn). Je interpreteert de ander dus als iemand die in zijn doen en laten betrokken is op de wereld zoals jij die zelf ook ervaart en inschat.

Bekijk nogmaals bovenstaand voorbeeld: Je vriendin is op weg naar haar moeder om haar bij te staan omdat zij plotseling ernstig ziek is geworden. Deze verklaring veronderstelt dat het *waar* is dat de moeder van je vriendin plotseling ernstig ziek is geworden. Het klinkt absurd als je zou zeggen: "V is op weg naar haar moeder omdat haar moeder plotseling ernstig ziek is geworden. Maar haar moeder is niet ziek." De verklaring van V's gedrag in het voorbeeld is 'factief'; door de verklaring te accepteren committeer je je er *zelf* aan dat het een feit is dat de moeder van V plotseling ernstig ziek is geworden. Contrasteer dit met een verklaring in termen van belief: V is op weg naar haar moeder omdat ze *denkt* dat haar moeder plotseling ernstig ziek is geworden. Maar haar moeder is niet ziek." Een dergelijke uitspraak is allesbehalve absurd. En dat komt doordat je hier expliciet onderscheid maakt tussen V's subjectieve representatie van de gezondheidstoestand van haar moeder en de haar werkelijke (als werkelijk veronderstelde) gezondheidstoestand.

Representational mindreading projecteert een subjectieve representatie van de wereld op de wereld zoals je hem zelf ervaart, een soort privé leefwereld van de ander die onderscheiden wordt van de wereld waarin je zijn gedrag plaatst. Relational mindreading daarentegen veronderstelt een *gedeelde, publieke* leefwereld waarin de ander wordt gerelateerd aan de doelen en redenen die oplichten in zijn praktische situatie. Het Relationele Model stelt dat deze relationele, publieke conceptie van mind aan de basis ligt van folk psychology. Verklaringen van andermans gedrag in termen van gebeurtenissen, feiten of standen van zaken in de buitenwereld zijn mentaliserend, niet doordat ze mentale representaties van een 'mindless' buitenwereld veronderstellen, maar doordat deze gebeurtenissen, feiten en standen van zaken impliciet als 'mindful' worden beschouwd, d.w.z. als wezenlijke onderdelen van de intentionele relaties van andermans handelen. Een paar pagina's terug omschreef ik mentaliseren als het vermogen om iemands gedrag te interpreteren in de context van zijn of haar mentale leefwereld. Het Relationele Model laat zien dat deze mentale leefwereld niets meer en niets minder hoeft te zijn dan de wereld waarin we leven.

De conceptuele uitdaging voor dit proefschrift is om deze relationele conceptie van mind hard te maken. Dit heeft echter alleen kans van slagen als we ons bevrijden van het mind-wereld dualisme dat inherent is aan het BD-Model. Hiertoe bespreek ik in hoofdstuk 3 een beroemd gedachte-experiment van de

filosoof Wilfrid Sellars uit 1956, dat aan de basis stond van het filosofische debat over folk psychology. Sellars' gedachte-experiment wordt vaak beschouwd als een vorm van conceptuele analyse van ons representationele begrip van mind. Ik laat zien waarom dit niet klopt en waarom het juist een perfecte manier is om een *relationele* conceptie van discursieve interactie te introduceren. In hoofdstuk 4 worden de relationele tegenhangers van beliefs en desires verder uitgewerkt tegen de achtergrond van dominante verklarende theorieën in het debat: de verschillende versies van de theorie theorie en de simulatie theorie. Ik laat zien dat al deze verklarende theorieën in wezen compatibel zijn met een relationele conceptie van het *explanandum* van discursieve sociale cognitie.

Hoofdstuk 5 laat vervolgens zien waarom het Relationele Model inderdaad een vruchtbaarder karakterisering geeft van doel-redeninterpretatie dan het BD-Model. Het Relationele Model kan uitleggen hoe we in staat zijn elkaar vliegensvlug te begrijpen in alledaagse situaties, het maakt inzichtelijk hoe we anderen *kennis* kunnen toeschrijven (i.p.v. slechts overtuigingen) en het geeft antwoord op de vraag hoe kinderen op jonge leeftijd geïntroduceerd worden in de wereld van folk psychology. Op al deze punten scoort het BD-Model aanzienlijk slechter. Hoofdstuk 6 gaat in op de additionele functie van representational mindreading. De praktische waarde van belief-desire ascriptie ligt in de mogelijkheden die het schept om om te gaan met anderen wanneer ze de normen lijken te schenden, om verschillen van mening uit te praten, om consensus te creëren, of om kritisch te reflecteren op onze normatieve praktijken zelf, bijvoorbeeld wanneer we vinden dat ze verbeterd kunnen worden.

Hoofdstukken 1 en 7 vormen samen het sluitstuk van dit proefschrift. Ze fungeren niet alleen als introductie op, resp. samenvatting van de andere hoofdstukken, maar geven tegelijkertijd antwoord op een achterliggende vraag: hoe komt het dat de relationele vorm van discursieve sociale cognitie nagenoeg onopgemerkt is gebleven in de filosofische discussies over folk psychology? Het antwoord dat in hoofdstuk 1 geponeerd en in hoofdstuk 7 beargumenteerd wordt, is dat we als filosofen vatbaar zijn voor een *reflectieve denkfout*. Wanneer filosofen reflecteren op alledaagse sociale cognitie, zijn ze geneigd hun eigen kritische, reflectieve interpretatie van de sociale situatie te projecteren op ons alledaagse begrip ervan. Filosofische reflectie is bij uitstek een cognitieve exercitie waarbij belief-desire psychologie goed van pas komt, zelfs onontbeerlijk is. Vanuit een filosofische houding lijkt het alsof de begrippen van doelen en redenen voor handelen, de begrippen van desire en belief veronderstellen. Als het Relationele Model klopt, is dit echter slechts kritische schijn. De filosofische praktijk is een uitzonderlijke sociale praktijk. Het kan

alleen maar vertekenend werken als we de psychologie van alledaagse sociale interactie modelleren naar de reflectieve houding van iemand die deze interactie slechts van een afstand aanschouwt.

Curriculum Vitae

Derek Willem Strijbos was born in Nijmegen on August 24th, 1979. He attended high school (Gymnasium) at Bernardinuscollege in Heerlen, where he graduated *cum laude* in 1997. He returned to Nijmegen to study medicine, which he combined with philosophy from 2002 onwards. After earning his MD in 2004, he worked at an addiction treatment center for 3 years, completing his MA-thesis in philosophy *cum laude* in 2006. He joined the department of philosophy of the Radboud University Nijmegen as a junior researcher (PhD-student) in 2007, where he worked on his PhD-thesis under supervision of prof. dr. Slors, published papers on social cognition and folk psychology, and taught several courses on the philosophy of mind and the philosophy of psychology. In the fall of 2011, Derek started his residency training in psychiatry at Dimence in Zwolle under supervision of prof. dr. Glas. He has a part-time research position at the Radboud University, currently focusing his research on the philosophy of psychiatry. Derek is married to Ayse Baltacı. Together they have two sons, Süleyman and Isaak.

Author Index

Understanding people in terms of their goals and reasons lies at the heart of our socio-cognitive abilities. It reveals the human mind as a discursive mind, a mind capable of reasoning about what to do and why. Established wisdom in many corners of philosophy and psychology has it that discursive social practice hinges on the ascription of representational mental states, prototypically beliefs and desires. On this so-called belief-desire model, we interpret the discursive minds of others essentially and exclusively as representational minds. This dissertation departs from this consensus view and argues that our quotidian, spontaneous understanding of one another in terms of goals and reasons is a form of relational sense-making. Accordingly, folk psychological practice is based on the ascription of relational mental states, states that relate the individuals we interact with to their goals and reasons out in the public world of common understanding. On this account, representational belief-desire psychology is not the conceptual core of folk psychology, nor the driving psychological force behind our discursive engagements with one another. Rather, it is an essentially complementary form of social cognition, designed for reflection on and management of such discursive engagements when our default, relational modus of understanding runs aground.

ISBN/EAN: 9789090271194